Nicholas J. Gotelli
University of Vermont

Aaron M. Ellison
Harvard Forest

PRINCÍPIOS DE ESTATÍSTICA EM ECOLOGIA

Consultoria, supervisão e revisão técnica desta edição:

Victor Lemes Landeiro
Biólogo. Mestre em Ciências Biológicas (Entomologia) pelo Instituto Nacional de Pesquisas da Amazônia (INPA).
Doutorando em Ciências Biológicas (Ecologia) pelo INPA.

Reimpressão 2011

artmed

2011

Obra originalmente publicada sob o título
A primer of ecological statistics
ISBN 978-0-87893-269-6
Copyright © 2004 by Sinauer Associates, Inc. All rights reserved.
Tradução autorizada por Sinauer Associates, Inc., Sunderland, MA 01375, USA.

Capa: *Mário Röhnelt*

Arte da capa: *copyright © 2004 Elizabeth Farnsworth*

Preparação de originais: *Greice Zenker Peixoto*

Leitura final: *Débora Benke de Bittencourt, Rebeca dos Santos Borges*

Editora sênior – Biociências: *Letícia Bispo de Lima*

Projeto e editoração: *Techbooks*

Para
Maryanne
&
Elizabeth

Reservados todos os direitos de publicação, em língua portuguesa, à
ARTMED® EDITORA S.A.
Av. Jerônimo de Ornelas, 670 – Santana
90040-340 – Porto Alegre – RS
Fone: (51) 3027-7000 Fax: (51) 3027-7070

É proibida a duplicação ou reprodução deste volume, no todo ou em parte, sob quaisquer formas ou por quaisquer meios (eletrônico, mecânico, gravação, fotocópia, distribuição na Web e outros), sem permissão expressa da Editora.

Unidade São Paulo
Av. Embaixador Macedo Soares, 10.735 – Pavilhão 5 – Cond. Espace Center
Vila Anastácio – 05095-035 – São Paulo – SP
Fone: (11) 3665-1100 Fax: (11) 3667-1333

SAC 0800 703-3444

IMPRESSO NO BRASIL
PRINTED IN BRAZIL

Quem mensura o paraíso, a terra, o mar e o céu
De modo a buscar folclore ou excitação
Deixe-o avisado de um idiota ser.

Sebastian Brant, *Ship of Fools*, 1494. Basel, Suíça

[N]úmeros são palavras sem as quais a exata descrição de qualquer fenômeno natural é impossível. ... Seguramente, qualquer fenômeno objetivo, de qualquer tipo, é quantitativo, bem como qualitativo; ignorar o primeiro, ou deixá-lo de lado como desimportante, virtualmente substitui a natureza objetiva por brinquedos abstratos, totalmente desprovidos de dimensões – brinquedos que nem existem nem podem ser concebidos a existir.

E. L. Michael, "Marine ecology and the coefficient of association:
A plea in behalf of quantitative biology", 1920.
Journal of ecology 8: 54-59.

[N]ós agora sabemos que aquilo que chamamos de leis da natureza são meramente verdades estatísticas e, desta forma, devem necessariamente permitir exceções. ... [N]ós precisamos do laboratório com suas restrições incisivas com objetivo de demonstrar a validade invariável da lei da natureza. Se deixarmos as coisas para a natureza, veremos uma cena bastante diferente: qualquer processo é parcial ou totalmente interferido pelo acaso, tanto assim que, sob circunstâncias naturais, o curso dos eventos, absolutamente em conformidade com leis específicas, é quase uma exceção.

Carl Jung, Foreword to *The I Ching or Book of Changes*.
Third Edition, 1950, traduzida por R. Wilhelm e C. F. Baynes.
Bollingen Series XIX, Princeton University Press.

Apresentação à Edição Brasileira

Um dos aspectos mais importantes da revolução científica nos séculos XVII e XVIII foi a adoção de um pensamento quantitativo (matemático) na ciência. Isso permitiu o desenvolvimento de modelos quantitativos capazes de fazer predições sobre os fenômenos naturais, além da confrontação dessas predições com dados empíricos. Esta última etapa passou a ser particularmente importante a partir do desenvolvimento da estatística, já no século XX. Nessa época, as diferentes áreas da biologia, considerando sua grande heterogeneidade, seguiram caminhos divergentes. Algumas áreas adotaram rapidamente uma visão experimental e quantitativa, seguindo a tradição mecanicista da física e da química, enquanto outras continuaram com uma visão mais descritiva e naturalística, ainda muito associada a um pensamento metafísico e vitalista, sobretudo pela falta de uma grande teoria unificadora que permitisse compreender a complexidade dos sistemas vivos.

Seria de se esperar, portanto, que essa dicotomia desaparecesse depois do estabelecimento do darwinismo e da teoria sintética da evolução, entre os anos 30 e 40 do século XX. De fato, a consolidação do pensamento evolutivo se desenvolveu principalmente a partir dos trabalhos em genética populacional e genética evolutiva de Ronald A. Fisher, J. B. S. Haldane e Sewall Wright, com um fortíssimo arcabouço matemático e estatístico. A ecologia, por sua vez, sendo uma área relativamente nova da biologia, emergiu já no século XX em um contexto mais quantitativo desde o seu início, com uma visão fortemente quantitativa e experimental, partindo primeiro de modelos demográficos e ecossistêmicos e, logo após, associando processos evolutivos e populacionais no contexto da ecologia "evolutiva", desenvolvida sobretudo por E. Hutchinson, R. MacArthur, E. O. Wilson, dentre tantos outros.

Entretanto, as mudanças não ocorreram como esperado, e uma tradição puramente descritiva e naturalística ainda permanece forte dentro de várias áreas da biologia (incluindo alguns campos da ecologia). Em pleno século XXI, deveria estar claro que uma visão quantitativa de mundo é absolutamente imperativa para o desenvolvimento da ciência moderna, e que a falta dessa visão traz sérias consequências para a formação dos biólogos como cientistas, reduzindo sua capacidade de atuar de forma plena na resolução de questões importantes (e de fato emergenciais) da atualidade. O que ocorre, nesse contexto, é que muitos dos alunos de graduação em Ciências Biológicas iniciam o curso com barreiras de aprendizado, erigidas ainda em etapas anteriores da educação formal. E, durante a graduação, essas barreiras tendem a se intensificar, as disciplinas quantitativas passam a ser consideradas, infelizmente, difíceis, sem importância para a biologia e de pouca utilidade prática. Além disso, é importante destacar que, em muitos casos, os professores responsáveis por algumas dessas disciplinas, infelizmente, contribuem para a formação dessa barreira.

Minha experiência de 20 anos de ensino de bioestatística, modelagem e métodos quantitativos aplicados à ecologia e à biologia evolutiva mostra que não é fácil, para a maior parte dos alunos, romper essas barreiras de aprendizado. Por outro lado, quando isso acontece, esses alunos passam a se desenvolver academicamente de forma muito mais adequada e se tornam profissionais muito competentes, independente de sua área de atuação dentro da biologia.

Isso mostra, portanto, que o esforço dos professores em tentar ajudar os alunos a romper essas barreiras vale a pena. Para isso, é crucial ter, como ponto de partida, bons livros-texto. Sem dúvida, em um contexto de métodos quantitativos aplicados à biologia e à ecologia, há textos excelentes (destaco os meus preferidos aqui: *Biometry*, de R. R. Sokal e F. J. Rohlf, e *Numerical ecology*, de P. Legendre & E. Legendre), mas em inglês, o que torna difícil sua indicação porque muitos estudantes de graduação e pós não dominam a língua.

Princípios de estatística em ecologia, do Dr. Nicholas Gotelli, da Universidade de Vermont, e de Aaron Ellison, publicado agora em português pela Artmed, com certeza será importante para preencher a lacuna de livros-texto adequados para o ensino de bioestatística, sobretudo em um contexto de ecologia e biologia evolutiva. Um dos livros anteriores do Dr. Gotelli, *Primer of ecology**, também possui a mesma característica, sendo um excelente texto com uma visão mais quantitativa, auxiliando o ensino de ecologia ao dar aos estudantes a possibilidade de compreender, de modo mais detalhado, alguns modelos e conceitos, abordados de forma geral em textos mais básicos, como os de R. E. Ricklefs, M. Begon e E. Pianka.

Princípios de estatística em ecologia, de Gotelli & Ellison, possui todas as características importantes de um excelente livro-texto, escrito em uma linguagem atual e moderna, para o ensino de métodos quantitativos em ecologia, tanto para cursos básicos de bioestatística para a graduação em biologia e áreas afins como para cursos mais avançados em ecologia quantitativa. Embora alguns exemplos trabalhados sejam apresentados, a maior ênfase é dada na interpretação dos resultados das análises e na discussão de como diferentes problemas podem ser resolvidos utilizando-se métodos observacionais ou experimentais (o que o torna um pouco diferente das obras antes citadas e de vários outros textos de bioestatística em português, que tendem a enfatizar inclusive uma abordagem experimental). Creio que isso reflete a mudança na facilidade computacional com que os diferentes métodos podem ser aplicados atualmente e, ao mesmo tempo, um aumento enorme na sua complexidade (de modo que seria extremamente complicado, e de fato de pouca utilidade, apresentar exemplos detalhados).

Além do cuidado com a interpretação e aplicação dos métodos, existem, sem dúvida, diversos outros aspectos que merecem destaque nesta obra de Gotelli & Ellison e que a tornam muito atraente. Um deles é a apresentação clara e imparcial de diferentes arcabouços teóricos dentro da estatística, incluindo uma discussão sobre métodos clássicos de inferência, métodos bayesianos e processos de Monte Carlo. Outro aspecto importantíssimo é o resgate, na forma de notas contendo curtas biografias e curiosidades, do papel de cientistas importantes e do contexto histórico com que métodos e conceitos importantes foram desenvolvidos. O livro também tem uma seção sobre métodos multivariados, algo bastante importante para os ecólogos (mas, em geral, negligenciado nos textos básicos de bioestatística).

Enfim, é extremamente recompensador poder usar esta obra em nossos cursos de ecologia quantitativa, bioestatística e modelagem: Dr. Gotelli é, sem dúvida, um dos mais renomados pesquisadores em ecologia quantitativa da atualidade (e, além disso, um músico entusiasta – vale a pena conferir tudo isso em sua webpage, http://www.uvm.edu/~ngotelli/homepage.html). Este livro, que a Artmed nos proporciona agora em português, mostra claramente que, além de ser um pesquisador brilhante e inovador em vários aspectos, Dr. Gotelli possui uma preocupação genuína com a formação dos jovens pesquisadores interessados em métodos computacionais e estatísticos aplicados à ecologia.

Jose Alexandre Felizola Diniz Filho
Professor titular do Departamento de Ecologia do
Instituto de Ciências Biológicas da Universidade Federal de Goiás

*N. de R. T. Publicado no Brasil, em 2007, pela Editora Planta, sob o título *Ecologia*.

Prefácio

Por que mais um livro de estatística? Este campo já tem muitos textos de biometria, teoria de probabilidades, delineamento experimental e análise de variância. A resposta é que ainda é preciso encontrar um único texto que seja capaz de unir as duas necessidades específicas dos ecólogos e cientistas da área ambiental. A primeira é uma introdução geral à teoria de probabilidades, incluindo os pressupostos da inferência estatística e o teste de hipóteses. A segunda é uma descrição detalhada de delineamentos e análises específicas, tipicamente encontrados em ecologia.

Esta obra é resultado de várias décadas de ensino de estatística para estudantes de graduação e pós-graduação na University of Oklahoma, na University of Vermont, no Mount Holyoke College e na Harvard Forest. Representa nossa perspectiva sobre o que é importante na interface ecologia e estatística, e a forma mais efetiva de ensiná-la. Este é o livro que gostaríamos de ter em mãos quando estudamos ecologia na faculdade.

"Princípios"* sugere um livro curto, com mensagens simples e recomendações seguras para os usuários. Se a estatística fosse tão simples! O assunto é amplo, e novos métodos e programas de computador surgem para os ecólogos quase mensalmente. Apesar do tamanho do livro (algo para o qual ficamos atentos), ainda o vemos como uma destilação do conhecimento estatístico. Incluímos as exposições clássicas, bem como as discussões sobre os métodos e as técnicas mais recentes, com o objetivo de que seja útil tanto para graduandos como para pesquisadores experientes. Como na arte e na música, há um forte elemento de estilo e preferências pessoais na escolha e aplicação da estatística. Explicamos nossas predileções por certos tipos de análises e delineamentos, mas os leitores e usuários de estatística deverão, ao final, escolher seu próprio conjunto de ferramentas estatísticas.

COMO UTILIZAR ESTE LIVRO

Contrastando com muitas obras de estatística direcionadas a biólogos e ecólogos – textos de "biometria" –, este não é um livro de "receitas", nem é um conjunto de exercícios e problemas atrelados a um programa de computador em particular, que provavelmente estaria ultrapassado. Apesar de o livro conter equações e derivações, não se trata de um texto estatístico formal, com provas matemáticas detalhadas. É, sim, uma exposição de tópicos básicos e avançados importantes especificamente para ecólogos e cientistas da área ambiental. É indicado para ser lido e utilizado talvez como um livro suplementar a outro mais

* N. de R.T. No original, *A primer of ecological statistics*. A tradução literal de *primer* seria "cartilha", porém este termo é mais utilizado em literatura infantil na língua portuguesa, pelo menos no Brasil. Por isso, preferimos utilizar como título *Princípios de estatística em ecologia*, em vez de *Cartilha de estatística aplicada à ecologia*.

tradicional, ou como um livro independente para estudantes que tiveram uma introdução mínima à estatística. Esperamos, também, que esta obra tenha seu lugar na prateleira (ou no chão) de profissionais da área ambiental, que necessitam usar e interpretar estatística diariamente, mas que não tenham um treinamento formal ou o auxílio de estatísticos.

Aqui, fazemos bom uso de notas de rodapé. Tal recurso permite-nos expandir o conteúdo e tratar de temas mais complexos, que poderiam comprometer o texto principal. Algumas notas de rodapé são puramente históricas; outras cobrem detalhes de provas matemáticas e estatísticas; outras, ainda, são breves ensaios sobre tópicos da literatura em ecologia. Quando graduandos, éramos apaixonados pelo clássico de Hutchinson, *An introduction to population biology* (Yale University Press 1977). O uso liberal de notas de rodapé por Hutchinson e suas frequentes incursões na história e filosofia nos serviram de modelo estilístico.

Encontrar um equilíbrio entre concisão e abrangência também esteve entre nossos objetivos. Muitos tópicos aparecem em mais de um capítulo, apesar de algumas vezes isso ocorrer em contextos um pouco diferentes. Como o Capítulo 12 requer álgebra de matrizes, providenciamos um breve apêndice que cobre o essencial em notação e manipulação de matrizes. Por fim, incluímos um glossário abrangente de termos usados em estatística e teoria de probabilidades, o qual pode ser útil também para a compreensão de métodos estatísticos apresentados em artigos científicos.

CONTEÚDO MATEMÁTICO

Apesar de não nos esquivarmos das equações quando necessárias, há consideravelmente menos matemática aqui do que em muitos textos de nível intermediário. Também tentamos ilustrar todos os métodos com análises empíricas. Em quase todos os casos, foram usados dados de nossos próprios estudos para ilustrar a progressão de dados brutos para resultados estatísticos.

Embora os capítulos sejam individualizados, há uma considerável referência cruzada. Os Capítulos 9-12 cobrem os tópicos tradicionais de muitos textos de biometria e contêm maior concentração de fórmulas e equações. O Capítulo 12, sobre métodos multivariados, é o mais avançado. Apesar de tentarmos explicá-los com uma linguagem simples (não foi uma tarefa fácil!), não há como "embelezar" a pesada dose de álgebra de matrizes e o novo vocabulário necessário para as análises multivariadas.

ABRANGÊNCIA E TÓPICOS

A estatística é um campo envolvente, no qual novos métodos sempre são desenvolvidos para responder questões ecológicas. Este livro reúne, contudo, o que acreditamos ser fundamental para qualquer análise estatística. O livro é organizado em torno de estatísticas com base para delineamento – para serem usadas em estudos observacionais ou experimentos delineados. (Uma estrutura alternativa seria estatística com base para modelos, incluindo critérios de seleção de modelos, análises de verossimilhança, modelagem inversa e outros métodos do tipo.)

Qual é a diferença entre essas duas abordagens? Em resumo, análises com base em delineamento tratam primeiramente o problema P (dados | modelo): acessar a probabilidade dos dados conforme o modelo que especificamos. Um elemento importante de análises com base em delineamento é ser explícito a respeito do modelo subjacente. Em contraste, estatísticas com base em modelos tratam o problema P (modelo | dados), ou seja, acessar a

probabilidade de um modelo conforme o conjunto de dados presente. Métodos com base em modelos frequentemente são mais apropriados para grandes conjuntos de dados que podem não ter sido coletados de acordo com uma estratégia amostral explícita.

Discutimos, também, modelos bayesianos, que de alguma forma estão entre essas duas abordagens. Métodos bayesianos tratam *P* (modelo | dados), mas podem ser usados com dados de experimentos e estudos observacionais delineados.

Esta obra foi dividida da seguinte forma: A Parte I discute os fundamentos de probabilidade e do pensamento estatístico, introduz a lógica e a linguagem de probabilidades (Capítulo 1), explica distribuições estatísticas comumente usadas em ecologia (Capítulo 2) e apresenta importantes medidas de alocação e dispersão (Capítulo 3). O Capítulo 4 mais se assemelha a uma discussão filosófica acerca de testes de hipóteses, com cuidadosas explicações dos erros Tipo I e Tipo II e a ubíqua premissa do "$P < 0,05$". O Capítulo 5 apresenta os três maiores paradigmas das análises estatísticas (frequentista, Monte Carlo e bayesiano), ilustrando seus usos através da análise de um único conjunto de dados.

A Parte II discute como delinear e executar estudos observacionais e experimentos de campo com sucesso. Seus capítulos reúnem conselhos e informações para serem utilizados antes da coleta de qualquer dado no campo. O Capítulo 6 trata dos problemas práticos de articular a razão para um estudo, ajustando um tamanho amostral e tratando a independência, a replicação, a aleatorização, os estudos de impacto e a heterogeneidade ambiental. O Capítulo 7 é um bestiário* em delineamento experimental. Contudo, em contraste com quase todos os outros capítulos, quase não há equações nele. Discutimos as forças e fraquezas de diferentes delineamentos sem encalhar em equações ou tabelas de análise de variância (ANOVA), que foram, em sua maioria, colocadas em quarentena na Parte III. O Capítulo 8 trata do problema de como manter e gerir dados uma vez coletados. Transformações de dados são introduzidas nesse capítulo como uma forma de ajudar na procura de pontos discrepantes (*outliers*) e erros na transcrição de dados. Também enfatizamos a importância de criar metadados – documentação que acompanha os dados brutos para uma armazenagem de longa duração.

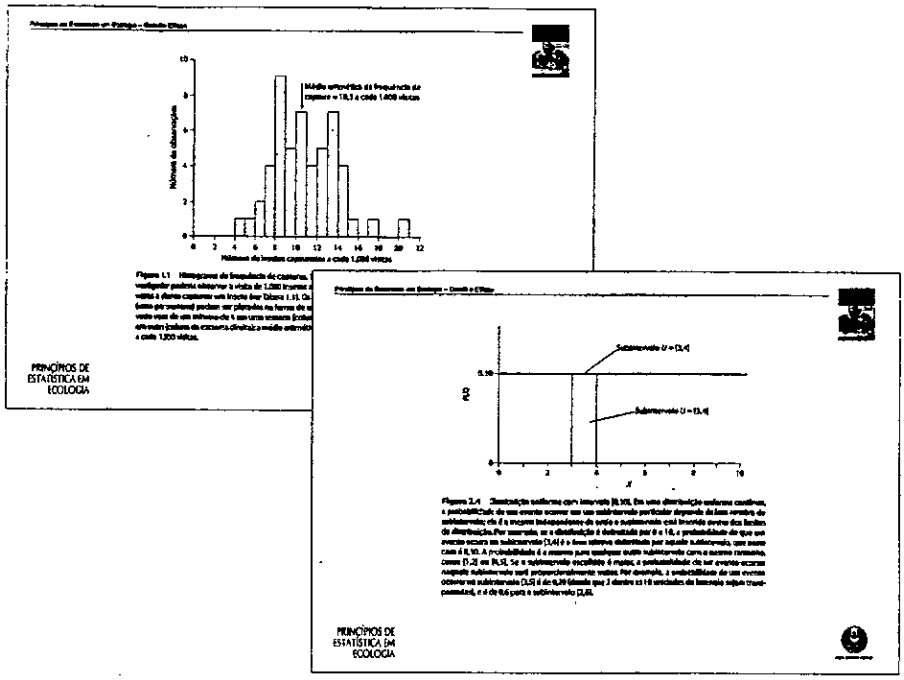
A Parte III discute análises específicas e engloba o conteúdo que é a essência da maioria dos textos de estatística. Seus capítulos desenvolvem muitos modelos básicos, mas também tentam apresentar mais algumas ferramentas atuais consideradas úteis pelos ecólogos e cientistas da área ambiental. O Capítulo 9 desenvolve o modelo básico de regressão linear, regressão de quantil, regressão robusta e análise de caminhos. O Capítulo 10 discute modelos de ANOVA (Análise de Variância) e de ANCOVA (Análise de Covariância), destacando qual modelo de ANOVA é apropriado para determinados delineamentos experimentais e amostrais. O capítulo também explica como usar contrastes *a priori* e enfatiza a representação gráfica dos resultados de ANOVA de um modo que facilita o entendimento dos efeitos principais e os termos de interação. O Capítulo 11 discute análise de dados categóricos, com ênfase em análises de chi-quadrado de tabelas de contingência de um e dois fatores e modelagem log-linear de tabelas de contingência múltiplas. O Capítulo 12 introduz uma variedade de métodos multivariados para ordenação e classificação. O capítulo apresenta, por fim, uma introdução à análise de redundância, um análogo multivariado da regressão múltipla.

---

*N. de R.T. Bestiário é um tipo de livro escrito na Idade Média que continha ilustrações e descrições de animais reais e imaginários com intuito de ensinar moral e entreter. No contexto deste livro, um "bestiário de delineamentos experimentais" indica que o capítulo será composto por figuras que ilustram cada tipo de delineamento experimental em adição ao texto que os descrevem.

RECURSOS DIDÁTICOS PARA PROFESSORES

Em www.artmed.com.br, professores terão acesso a PowerPoints® (em português) com as figuras da obra, que poderão ser utilizados para preparar suas aulas.



SOBRE A CAPA

A capa é uma pintura de natureza morta feita pela ecóloga Elizabeth Farnsworth, inspirada pelas pinturas de natureza morta holandesas do século XVII. Os pintores da Holanda, incluindo De Heem, Claesz e Kalf, mostraram sua habilidade arranjando e pintando uma grande variedade de objetos naturais e artificiais, frequentemente com ricas alusões e simbolismo oculto. O belo e biologicamente acurado desenho de plantas e animais de Elizabeth segue a longa tradição de ilustradores da biologia, como Dürer e Escher. No desenho da ecóloga, a peça central é uma ilustração do Teorema de Bayes feito à mão, uma exposição matemática do século XVIII que permeia a essência da teoria de probabilidades. Apesar de os métodos bayesianos modernos serem intensivos computacionalmente, o entendimento do Teorema de Bayes ainda requer que seja contemplado à mão.

As pinturas holandesas de natureza morta frequentemente incluem instrumentos musicais, talvez simbolizando a beleza pitagoreana da música e dos números. O desenho de Elizabeth incorpora o bouzouki (instrumento de cordas) de Aaron, como a escrita deste livro foi entrelaçada a muitas sessões de música acústica com Aaron e Elizabeth.

O crânio humano é o símbolo tradicional da mortalidade em pinturas de natureza morta. Elizabeth o substituiu por um pequeno coral cérebro, em homenagem ao primeiro trabalho de dissertação de Nick (Gotelli) sobre ecologia de corais *gorgonians* subtropicais. À luz do atual colapso mundial dos recifes de corais, o coral cérebro parece um símbolo

apropriado para mortalidade. Em primeiro plano, uma formiga *Atta* carrega um fragmento de texto com o sinal de somatório, relembrando-nos do poder coletivo de formigas operárias individuais, pois a soma de suas atividades beneficia toda a colônia. Como nas pinturas medievais, a formiga é desenhada desproporcionalmente maior que os outros objetos, refletindo sua dominância e importância nos ecossistemas terrestres. No segundo plano está um espécime de *Sarracenia minor*, planta carnívora do sudeste dos EUA que é cultivada em estufa. Apesar de muito rara para manipulações experimentais em campo, nossos estudos em andamento sobre a *Sarracenia purpurea*, mais comum no nordeste dos EUA, podem nos ajudar a desenvolver estratégias efetivas de conservação para todas as espécies do gênero.

O elemento final neste desenho é um "jarro galinha", da Renascença Italiana. No verão de 2001, estávamos investigando plantas-jarro em pântanos remotos das Montanhas Adirondack. Ao final de um longo dia de campo, em um pequeno restaurante italiano no sul de Vermont, pedimos e consumimos, um tanto sem moderação, um "jarro galinha" contendo um litro do vinho Chianti. Ao fim da noite, tínhamos nos comprometido a escrever este livro, e "A galinha" – como sempre o chamamos – oficialmente saiu do ovo.

AGRADECIMENTOS

Agradecemos a vários colegas por suas importantes sugestões para diferentes capítulos: Marti Anderson (11, 12), Jim Brunt (8), George Cobb (9, 10), Elizabeth Farnsworth (1-5), Brian Inouye (6, 7), Pedro Peres-Neto (11, 12), Catherine Potvin (9, 10), Robert Rockwell (1-5), Derek Roff (proposta original do livro), David Skelly (proposta original do livro) e Steve Tilley (1-5). Discussões em grupo no laboratório em Harvard Forest e na University of Vermont também nos forneceram um retorno para muitas partes do livro. Créditos especiais vão para Henry Horn, que leu e criticou em detalhes todo o manuscrito. Agradecemos a Andy Sinauer, Chris Small, Carol Wigg, Bobbie Lewis e Susan McGlew, da Sinauer Associates, e, especialmente, Michele Ruschhaupt, do The Format Group, por transformar as páginas sem graça do texto em word em um livro elegante e agradável.

A National Science Foundation ofereceu suporte ao nosso estudo com as plantas-jarro, formigas, modelos nulos e manguezais, os quais aparecem em muitos exemplos ao longo do livro. A University of Vermont e a Harvard Forest, nossas respectivas instituições de trabalho, proveram ambientes de suporte para sua escrita.

Sumário

PARTE I
FUNDAMENTOS DA PROBABILIDADE E DO PENSAMENTO ESTATÍSTICO

1 Introdução à Probabilidade 23
2 Variáveis Aleatórias e Distribuições de Probabilidades 45
3 Estatísticas Descritivas: Medidas de Posição e Dispersão 75
4 Formulando e Testando Hipóteses 97
5 Três Estruturas Para Análises Estatísticas 125

PARTE II
DELINEAMENTO DE EXPERIMENTOS

6 Delineando Estudos de Campo com Sucesso 155
7 Um Bestiário de Delineamentos Experimentais e Amostrais 181
8 Gestão e Curadoria de Dados 225

PARTE III
ANÁLISE DE DADOS

9 Regressão 257
10 Análise de Variância 307
11 Análise de Dados Categóricos 367
12 Análise de Dados Multivariados 401

Apêndice Álgebra de Matrizes para Ecólogos 465
Glossário 477
Literatura Citada 499
Índice 511

Sumário Detalhado

Parte I
FUNDAMENTOS DA PROBABILIDADE E DO PENSAMENTO ESTATÍSTICO

1 Introdução à Probabilidade 23

O que é probabilidade? 24
Medindo probabilidades 24
A probabilidade de um evento único: captura de presas por plantas carnívoras 24
Estimando probabilidades por amostragem 27
Problemas na definição de probabilidade 29
A matemática da probabilidade 31
Definindo o espaço amostral 31
Eventos complexos e compartilhados: combinando probabilidades simples 33
Cálculo de probabilidades: paineirinhas e lagartas 35
Eventos complexos e compartilhados: regras combinatórias de conjuntos 38
Probabilidades condicionais 41
Teorema de Bayes 42
Resumo 44

2 Variáveis Aleatórias e Distribuições de Probabilidades 45

Variáveis aleatórias discretas 46
Variáveis aleatórias de Bernoulli 46
Exemplo de uma tentativa de Bernoulli 47
Muitas tentativas de Bernoulli = uma variável aleatória binomial 48
A distribuição binomial 51
Variáveis aleatórias de Poisson 54
Um exemplo de variável aleatória de Poisson: distribuição de uma planta rara 56
O valor esperado de uma variável aleatória discreta 59
A variância de uma variável aleatória discreta 59
Variáveis aleatórias contínuas 61
Variáveis aleatórias uniformes 62
O valor esperado de uma variável aleatória contínua 64
Variável aleatória normal 65
Propriedades úteis da distribuição normal 68
Outras variáveis aleatórias contínuas 70
O teorema do limite central 72
Resumo 74

3 Estatísticas Descritivas: Medidas de Posição e Dispersão 75

Medidas de posição 76
A média aritmética 76
Outras médias 78
Outras medidas de posição: a mediana e a moda 82
Quando usar cada medida de posição 83
Medidas de dispersão 83
A variância e o desvio-padrão 84
O erro-padrão da média 85
Assimetria (*skewness*), curtose e momento central 87
Quantis 89

Usando as medidas de dispersão 90
Algumas questões filosóficas em torno das estatísticas descritivas 91
Intervalos de confiança 92
Intervalos de confiança generalizados 94
Resumo 95

4 Formulando e Testando Hipóteses 97
Métodos científicos 98
Dedução e indução 98
A indução nos dias modernos: inferência bayesiana 101
O método hipotético-dedutivo 105
Testando hipóteses estatísticas 108
Hipótese estatística *versus* retorno à hipótese científica 108
Significância estatística e valores de *P* 109
Erros em testes de hipóteses 117
Estimativa de parâmetros e predição 121
Resumo 123

5 Três Estruturas Para Análises Estatísticas 125
Problema amostral 125
Análises de Monte Carlo 128
Passo 1: especificando o teste estatístico 129
Passo 2: criando a distribuição nula 129
Passo 3: decidindo sobre um teste uni ou bicaudal 130
Passo 4: calculando a probabilidade de cauda 132
Premissas do método de Monte Carlo 133
Vantagens e desvantagens do método de Monte Carlo 133
Análises paramétricas 135
Passo 1: especificando a estatística-teste 135
Passo 2: especificando a distribuição nula 137
Passo 3: calculando a probabilidade de cauda 137
Premissas do método paramétrico 138
Vantagens e desvantagens do método paramétrico 139
Análises não paramétricas: um caso especial de análise de Monte Carlo 139
Análises bayesianas 140
Passo 1: especificando a hipótese 140
Passo 2: especificando os parâmetros como variáveis aleatórias 142
Passo 3: especificando a distribuição de probabilidades de priores 143
Passo 4: calculando a verossimilhança 147
Passo 5: calculando a distribuição de probabilidades *a posteriori* 147
Passo 6: interpretando os resultados 148
Premissas da análise bayesiana 150
Vantagens e desvantagens da análise bayesiana 150
Resumo 151

PARTE II
DELINEAMENTO DE EXPERIMENTOS

6 Delineando Estudos de Campo com Sucesso 155
Qual é a questão central do estudo? 155
Existem diferenças espaciais ou temporais na variável *Y*? 155
Qual é o efeito do fator *X* sobre a variável *Y*? 156
As medidas da variável *Y* são consistentes com as predições da hipótese H? 156
Usando as medidas da variável *Y*, qual é a melhor estimativa do parâmetro ϵ no modelo Z? 157
Experimentos manipulativos 157
Experimentos naturais 159
Experimentos fotográficos *versus* experimentos de trajetória 161
O problema da dependência temporal 162
Experimentos de pressão *versus* experimentos de pulsos 164
Replicação 166

Quantas réplicas? 166
Quantas réplicas são possíveis no total? 166
A Regra dos 10 167
Estudos em larga escala e impactos ambientais 168
Garantindo independência 169
Evitando fatores confundidos 171
Replicação e randomização 172
Delineando experimentos de campo e estudos amostrais efetivos 176
As parcelas ou os cercados são grandes o suficiente para garantir resultados realísticos? 176
Qual é o grão e a extensão do estudo? 176
A amplitude dos tratamentos ou das categorias dos censos engloba ou abrange a amplitude de condições ambientais possíveis? 177
Foram estabelecidos controles apropriados para garantir que os resultados refletem variação apenas no fator de interesse? 178
Todas as réplicas foram manipuladas da mesma forma, exceto para a aplicação intencional do tratamento? 178
As covariáveis apropriadas foram medidas em cada réplica? 179
Resumo 179

7 Um Bestiário de Delineamentos Experimentais e Amostrais 181

Variáveis categóricas *versus* variáveis contínuas 182
Variáveis dependentes e independentes 183
Quatro classes de delineamentos experimentais 183
Delineamentos de regressão 184
Delineamentos de ANOVA 189
Alternativas à ANOVA: experimento de regressão 215
Delineamentos tabulares 218
Alternativas para os delineamentos tabulares: delineamentos proporcionais 221
Resumo 222

8 Gestão e Curadoria de Dados 225

O primeiro passo: gestão de dados brutos 226
Planilhas 226
Metadados 227
O segundo passo: armazenamento e curadoria de dados 228
Armazenamento: temporário e permanente 228
Curadoria de dados 229
O terceiro passo: verificação dos dados 230
A importância dos dados discrepantes 230
Erros 232
Dados faltantes 233
Detectando discrepâncias e erros 233
Criando um rastro de auditoria 241
O passo final: transformação de dados 241
Transformações de dados como ferramenta cognitiva 242
Transformações de dados por demanda estatística 246
Apresentando resultados: transformados ou não? 250
O rastro de auditoria revisitado 251
Resumo: o fluxograma de gerenciamento de dados 251

PARTE III
ANÁLISE DE DADOS

9 Regressão 257

Definindo uma linha reta e seus dois parâmetros 257
Ajustando dados a um modelo linear 259
Variâncias e covariâncias 262

Estimativa de parâmetro por mínimos quadrados 264
Componentes de variância e coeficiente de determinação 266
Testes de hipóteses com regressão 268
A anatomia de uma tabela ANOVA 268
Outros testes e intervalos de confiança 271
Pressupostos da regressão 275
Testes de diagnóstico para a regressão 277
Representação gráfica dos resíduos 277
Outros gráficos de diagnóstico 280
A função de influência 280
Monte Carlo e análise bayesiana 282
Regressão linear usando métodos de Monte Carlo 283
Regressão linear usando métodos bayesianos 283
Outros tipos de análises de regressão 286
Regressão robusta 286
Regressão quantílica 289
Regressão logística 291
Regressão não linear 293
Regressão múltipla 293
Análise de caminhos 297
Critérios para seleção de modelos 300
Métodos de seleção de modelos para regressão múltipla 301
Métodos de seleção de modelos na análise de caminhos 302
Seleção de modelos bayesianos 303
Resumo 305

10 Análise de Variância 307
Símbolos e rótulos em ANOVA 308
Anova e partição da soma dos quadrados 308
Os pressupostos da ANOVA 313
Testes de hipóteses com ANOVA 314
Construindo razões-F 316
Um bestiário de tabelas de ANOVA 318
Blocos aleatorizados 318
ANOVA aninhada 320
ANOVA de dois fatores 322
ANOVA para delineamentos com três fatores e para *n*-fatores 326
ANOVA de parcelas subdivididas 326
ANOVA de medidas repetidas 327
ANCOVA 332
Fatores aleatórios *versus* fatores fixos em ANOVA 335
Partição de variância na ANOVA 340
Após a anova: plotando e entendendo os termos de interação 342
Plotando resultados de ANOVAs de um fator 343
Plotando resultados de ANOVAs de dois fatores 345
Entendendo os termos de interação 349
Plotando resultados de ANCOVAs 350
Comparando médias 352
Comparações *a posteriori* 354
Contrastes *a priori* 356
Correções de Bonferroni e o problema dos testes múltiplos 362
Resumo 366

11 Análise de Dados Categóricos 367
Tabelas de contingência de dois fatores 368
Organizando os dados 368
As variáveis são independentes 370
Teste de hipóteses: teste de chi-quadrado de Pearson 372
Uma alternativa ao teste de chi-quadrado de Pearson: o teste-*G* 375
O teste de chi-quadrado e o teste-*G* para tabelas $L \times C$ 377
Qual teste escolher? 381
Tabelas de contingência multifatoriais 382
Organizando os dados 382
Sobre as tabelas multifatoriais! 386
Abordagens bayesianas para tabelas de contingência 393
Testes para bondade do ajuste 393
Teste de bondade do ajuste para distribuições discretas 394

Teste da bondade do ajuste a distribuições contínuas: o teste de Kolmogorov-Smirnov 398
Resumo 399

12 Análise de Dados Multivariados 401

Abordando dados multivariados 402
A necessidade da álgebra de matrizes 402
Comparando médias multivariadas 405
Comparando médias multivariadas de duas amostras: Teste T^2 de Hotelling 405
Comparando médias multivariadas de mais de duas amostras: uma MANOVA simples 408
A distribuição normal multivariada 412
Teste de normalidade multivariada 413
Medidas de distância multivariada 416
Medindo distâncias entre dois indivíduos 416
Medindo distâncias entre dois grupos 420
Outras medidas de distância 420
Ordenação 424
Análise de componentes principais 424
Análise fatorial 433
Análise de coordenadas principais 436
Análise de correspondência 439
Escalonamento multidimensional não métrico 443
Vantagens e desvantagens da ordenação 445
Classificação 447
Análise de agrupamentos 447
Escolhendo um método de agrupamento 448
Análise discriminante 451
Vantagens e desvantagens da classificação 455
Regressão múltipla multivariada 455
Análise de redundância 456
Resumo 462

Apêndice Álgebra de Matrizes para Ecólogos 465

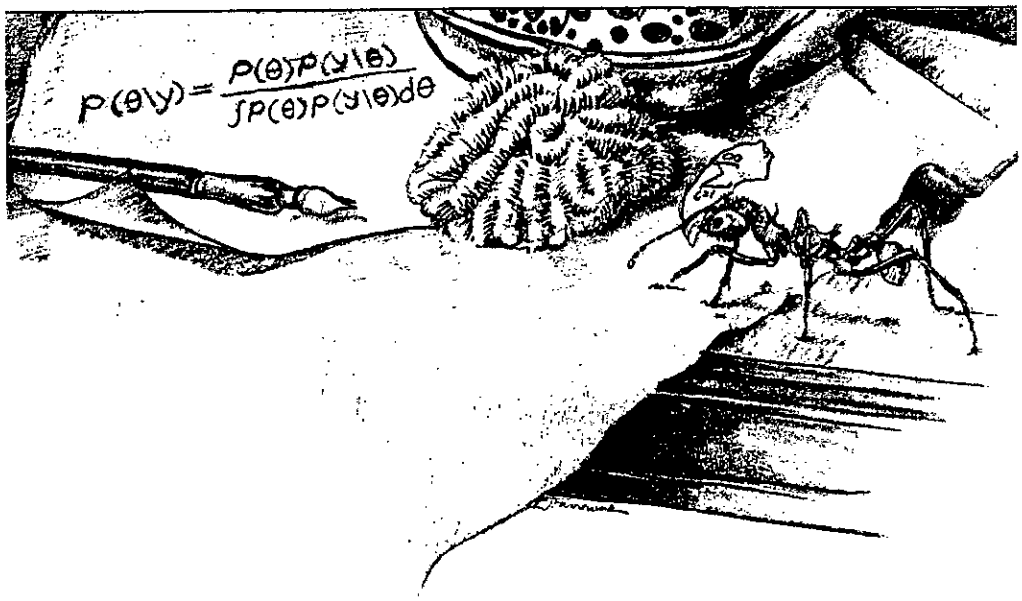
Glossário 477

Literatura Citada 499

Índice 511

PARTE I

FUNDAMENTOS DA PROBABILIDADE E DO PENSAMENTO ESTATÍSTICO



1

Introdução à Probabilidade

Neste capítulo, desenvolveremos conceitos básicos e definições necessárias para entender probabilidade e amostragem. Os detalhes do cálculo de probabilidades estão por trás da computação de todos os testes estatísticos de significância. Desenvolver uma apreciação pela probabilidade ajuda a melhor delinear experimentos e a interpretar seus resultados com mais clareza.

Os conceitos deste capítulo estabelecem os alicerces para o uso e entendimento da estatística. Eles são mais importantes, de fato, do que alguns dos tópicos detalhados que o seguirão. Este capítulo fornece o conhecimento necessário para questões críticas como "O que quer dizer, em um artigo científico, que as médias de duas amostras diferem significativamente a um $P = 0{,}003$?" ou "Qual a diferença entre os erros estatísticos do Tipo 1 e do Tipo 2?". Você ficaria surpreso ao saber que muitos cientistas com considerável experiência em estatística são incapazes de responder claramente a essas questões? Se você entender o conteúdo deste capítulo introdutório, terá uma base forte para o seu estudo de estatística, e estará sempre apto a interpretar corretamente o conteúdo estatístico na literatura científica – mesmo que não esteja familiarizado com os detalhes particulares dos testes utilizados.

Aqui, também apresentamos o problema de mensuração e quantificação, processos que são essenciais a todas as ciências. Não podemos começar a investigar fenômenos cientificamente, a menos que possamos quantificar processos e concordar, por meio de uma linguagem comum que possibilitará a interpretação de nossas medições. É claro que o mero ato de quantificação por si só não faz de algo uma ciência: a astrologia e a previsão do mercado de ações utilizam muitos números, mas não se qualificam como ciências.

Um desafio conceitual para estudantes de ecologia é traduzir seu "amor pela natureza" em um "amor ao padrão." Por exemplo, como quantificamos padrões na abundância de plantas e de animais? Quando caminhamos pela floresta, nos fazemos uma série de questões, como: "Qual a nossa melhor estimativa para a densidade de colônias de *Myrmica* (formiga comum em florestas) nesta floresta? Seria de 1 colônia/m^2? 10

colônias/m^2? Qual seria a melhor maneira de medir a densidade de *Myrmica*? Como a densidade de *Myrmica* varia em diferentes partes da floresta? Quais mecanismos ou hipóteses poderiam explicar tal variação?" E, por fim, "Quais experimentos e observações poderíamos fazer para estudar e testar ou refutar essas hipóteses?"

Porém, uma vez que "quantificamos a natureza", ainda temos de sumarizar, sintetizar e interpretar os dados que coletamos. Em todas as ciências, a estatística é a linguagem comum usada para interpretar nossas medidas e para testar e discriminar entre nossas hipóteses. A probabilidade é a base da estatística e, portanto, é o ponto de partida deste livro.

O QUE É PROBABILIDADE?

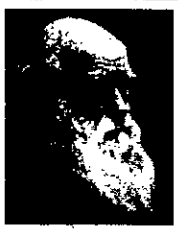
Se a previsão do tempo diz que há 70% de chance de chover, intuímos o que isso significa. Tal declaração quantifica a **probabilidade**, ou o provável resultado de um evento sobre o qual estamos inseguros. A incerteza existe porque há **variação** no mundo, a qual nem sempre é previsível. A variação em sistemas biológicos é especialmente importante; é impossível entender conceitos básicos de ecologia, evolução e ciências ambientais sem uma apreciação da variação natural.[1]

Apesar de todos termos uma noção geral de probabilidade, defini-la com precisão é outra questão. O problema em definir probabilidade é mais sério quando tentamos medir a de eventos reais.

MEDINDO PROBABILIDADES

A probabilidade de um evento único: captura de presas por plantas carnívoras

As plantas carnívoras são bons sistemas para se pensar acerca da definição de probabilidade. A planta-jarro do norte* *Sarracenia purpurea* captura insetos em folhas cheias de água da chuva.[2] Alguns dos que visitam uma planta-jarro



Charles Darwin

[1] A variação em peculiaridades entre indivíduos é um dos elementos-chave da teoria da evolução e da seleção natural. Uma das grandes contribuições intelectuais de Charles Darwin (1809-1882) foi enfatizar a significância de tal variação e de abandonar a camisa de força conceitual do "conceito tipológico de espécie" que tratava as espécies como entidades fixas e estáticas com fronteiras imutáveis e bem-definidas.

Planta-jarro (Sarracenia purpurea)

[2] Plantas carnívoras são um dos nossos organismos de estudos favoritos. Elas fascinam biólogos desde que Darwin (1875) provou pela primeira vez que elas podem absorver nutrientes dos insetos. Plantas carnívoras possuem muitos atributos que as fazem sistemas ecológicos modelo (Ellison & Gotelli, 2001; Ellison et al., 2003).

*N. de T. Em inglês, *northern pitcher plant* é uma planta com folhas que se assemelham a um jarro de água, por isso denominada "planta-jarro".

caem dentro dela e afundam; a planta extrai nutrientes da presa em decomposição. Apesar de a armadilha ser uma admirável adaptação evolucionária para a vida carnívora, ela não é tão eficiente, e a maioria dos insetos que visita a planta-jarro escapa.

Como poderíamos estimar a probabilidade de que a visita de um inseto irá resultar em uma captura de sucesso? A forma mais simples seria registrar cuidadosamente o número de insetos que visitam a planta e quantos são capturados. *Visitar a planta* é um exemplo do que os estatísticos chamam de **evento.** Um evento é um processo simples, de início e fim bem reconhecíveis.

Neste universo simples, visitar uma planta é um evento que pode ter dois **resultados**: *fuga* ou *captura*. A captura da presa é um exemplo de **resultado discreto,** porque a ele pode ser atribuído um número inteiro positivo. Por exemplo, você poderia atribuir 1 para uma captura e 2 para uma fuga. O **conjunto** formado por todos os resultados possíveis de um evento é chamado de **espaço amostral.**[3] Espaços amostrais ou conjuntos compostos de resultados discretos são chamados de **conjuntos discretos,** porque seus resultados são todos contáveis.[4]

[3] Mesmo nesse exemplo simples, as definições são traiçoeiras e não abrangem todas as possibilidades. Por exemplo, alguns insetos voadores podem pairar sobre a planta e nunca tocá-la, e alguns rastreadores, no entanto, podem explorar a superfície externa da planta, sem ultrapassar a borda, e entrar no jarro. Dependendo de quão preciso tentamos ser, essas possibilidades podem ou não ser incluídas no espaço amostral de observações usado para determinar a probabilidade de ser capturado. O espaço amostral estabelece o domínio de inferência do qual podemos tirar conclusões.

G. F. L. P. Cantor

[4] O termo *contável* tem um significado muito específico na matemática. Um conjunto é considerado contável se a todos os elementos (ou resultados) daquele conjunto pode ser atribuído um valor inteiro. Por exemplo, podemos atribuir aos elementos do conjunto *visita* os inteiros 1 (= captura) e 2 (= fuga). O conjunto de números inteiros por si só é contável, mesmo sabendo que há infinitos números inteiros. No final dos anos 1800, o matemático Georg Ferdinand Ludwig Philipp Cantor (1845-1918), um dos fundadores do que veio a ser chamado de teoria dos conjuntos, desenvolveu a noção de cardinalidade para descrever tal contabilidade. Dois conjuntos são considerados como tendo a mesma cardinalidade se eles podem ser colocados em correspondência um com o outro de um para um. Por exemplo, os elementos do conjunto de todos os números pares {2, 4, 6, ...} podem ser cada um atribuídos a um elemento do conjunto de todos os números inteiros (e vice-versa). Conjuntos que possuem a mesma cardinalidade que os números inteiros são ditos como tendo cardinalidade 0, e são denotados como $\aleph_0$. Cantor tentou provar que o conjunto de números racionais (os que podem ser escritos como o quociente entre quaisquer dois números inteiros) também possuem cardinalidade $\aleph_0$, mas o conjunto de números irracionais (os que não podem ser escritos como o quociente entre quaisquer dois números inteiros, como $\sqrt{2}$) não são contáveis. O conjunto de números irracionais tem cardinalidade 1 e é denotado $\aleph_1$. Um dos resultados mais famosos de Cantor foi de que existe uma correspondência de um para um entre o conjunto de todos os pontos do pequeno intervalo [0,1] e o conjunto de todos os pontos em um espaço *n*-dimensional (ambos têm cardinalidade 1). Tal resultado, publicado em 1877, o levou a escrever para o matemático Julius Wilhelm Richard Dedekind: "Eu vejo isso, mas eu não acredito nisso!". Curiosamente, existem tipos de conjuntos que possuem cardinalidade maior que 1, e de fato existem infinitamente muitas cardinalidades. Ainda estranho, o conjunto dos números cardinais ($\aleph_0$, $\aleph_1$, $\aleph_2$, ...) é, por si só, um conjunto contável infinito.

Cada inseto que visita a planta é contado como uma **observação***. Estatísticos em geral se referem a cada tentativa como uma **réplica** individual e a um conjunto de tentativas como um **experimento**.[5] Definimos a probabilidade de um resultado como o número de vezes que ele ocorre dividido pelo número de observações. Se olhássemos uma única planta e registrássemos 30 capturas de presas dentre 3.000 visitas, poderíamos calcular a probabilidade como

$$\frac{\text{número de capturas}}{\text{número de visitas}}$$

ou 30 em 3.000, que podemos escrever como 30/3.000, ou 0,01. Em termos mais gerais, calculamos a probabilidade P de um resultado ocorrer como sendo

$$P = \frac{\text{número de resultados}}{\text{número de observações}} \qquad (1.1)$$

Por definição, nunca pode haver mais resultados do que observações, então o numerador desta fração nunca é maior que o denominador, portanto

$$0{,}0 \leq P \leq 1{,}0$$

Em outras palavras, probabilidades são sempre restringidas a um mínimo de 0,0 e a um máximo de 1,0. Uma probabilidade de 0,0 representa um evento que nunca ocorrerá, uma probabilidade de 1,0, um evento que sempre ocorrerá.

Contudo, mesmo esta estimativa é problemática. A maioria das pessoas poderia dizer que a probabilidade de que o sol irá nascer amanhã é uma coisa certa (P = 1,0). Certamente, se registrássemos assiduamente o nascer do sol a cada manhã, verificaríamos que ele sempre nasce, e nossa medida poderia, de fato, fornecer uma probabilidade estimada de 1,0. Mas o sol é uma estrela ordinária, um reator nuclear que libera energia quando os átomos de hidrogênio se fundem para formar o hélio. Após cerca de 10 bilhões de anos, todo o combustível de hidrogênio se esgota e a estrela morre. O sol é uma estrela de meia idade, de forma que, se você pudesse continuar observando-o por 5 bilhões de anos, certa manhã o sol poderia não nascer mais. E se você iniciasse suas observações naquele ponto, sua estimativa da probabilidade do sol nascer seria de 0,0. Então, a probabilidade de o sol nascer a cada dia é realmente 1,0? Pouco menos que 1,0? É 1,0 na atualidade, mas 0,0 daqui a 5 bilhões de anos?

Este exemplo ilustra que qualquer estimativa ou medida de probabilidade condiz com a maneira como definimos o espaço amostral – o conjunto de

[5] Esta definição estatística de experimento é menos restritiva que a convencional, que descreve um conjunto de aspectos manipulados (o tratamento) e um grupo comparativo apropriado (o controle). Contudo, muitos ecólogos usam o termo **experimento natural** para se referir a comparações entre réplicas que não foram manipuladas pelo investigador, mas que diferem naturalmente na quantidade de interesse (p. ex., ilhas com e sem predadores; *ver* Capítulo 6).

* N. de T. Em inglês, *trial* também pode ser traduzido como "tentativa".

todos os resultados possíveis que usamos para comparação. Em geral, todos temos estimativas intuitivas ou palpites de probabilidades para todos os tipos de eventos do cotidiano. Contudo, para quantificar esses palpites, devemos decidir sobre um espaço amostral, colher amostras, e contar a frequência de certos eventos.

Estimando probabilidades por amostragem

Em nosso primeiro experimento, observamos a visita de 3.000 insetos à planta; 30 deles foram capturados e calculamos a probabilidade de captura de 001. Este número é razoável ou nosso experimento foi conduzido em um dia particularmente bom para as plantas (ou um dia ruim para os insetos)? Sempre haverá variabilidade neste tipo de números. Em alguns dias, nenhuma presa é capturada; em outros, há muitas capturas. Poderíamos determinar precisamente a verdadeira probabilidade de captura se observássemos cada inseto visitando uma planta todo tempo do dia ou da noite, a cada dia do ano. Cada vez que houvesse uma visita poderíamos determinar quando o resultado *Captura* ocorreu, e como consequência atualizar nossa estimativa da probabilidade de captura. Mas a vida é muito curta para isso. Como poderíamos estimar a probabilidade de captura sem monitorar as plantas de forma constante?

Podemos estimar de maneira eficiente a probabilidade de um evento coletando uma **amostra** da população de interesse. Por exemplo, em todas as semanas, por um ano inteiro, poderíamos escolher um dia e observar 1.000 visitas de insetos às plantas. O resultado é um conjunto de 52 amostras de 1.000 observações cada, e o número de insetos capturados em cada amostra. As três primeiras linhas desse conjunto de dados de 52 linhas são mostradas na Tabela 1.1. As probabilidades de captura diferem entre os diferentes dias, mas não por muito;

TABELA 1.1 Dados amostrais para o número de capturas de insetos a cada 1.000 visitas à planta carnívora *Sarracenia purpurea*

Número ID	Data da observação	Número de insetos capturados
1	1 de junho de 1998	10
2	8 de junho de 1998	13
3	15 de junho de 1998	12
⋮	⋮	⋮
52	24 de maio de 1999	11

Neste conjunto de dados hipotético, cada linha representa uma semana na qual uma amostra foi coletada. Se a amostragem foi conduzida uma única vez na semana por um ano inteiro, o conjunto de dados deve ter exatamente 52 linhas. A primeira coluna indica o número de identidade, uma sequência única de números inteiros (1 a 52) atribuídos a cada linha do conjunto de dados. A segunda coluna informa a data da amostra e a terceira o número de capturas de insetos dentre as 1.000 visitas observadas. A partir desse conjunto de dados, a probabilidade de captura para cada amostra pode ser estimada como o número de capturas/1.000. Os dados de todos os 52 números de captura podem ser plotados e sumarizados em um histograma que ilustra a variabilidade entre as amostras (*ver* Figura 1.1).

parece ser relativamente incomum para os insetos serem capturados enquanto visitam uma planta-jarro.

Na Figura 1.1, sumarizamos concisamente os resultados de um ano de amostras[6] em um gráfico chamado **histograma**. Neste, em particular, os números no eixo horizontal, ou **eixo-*x***, indicam quantos insetos foram capturados enquanto visitavam plantas. Os números variam de 4 a 20, porque, nas 52 amostras, houve dias em que apenas 4 foram capturados, dias em que 20 foram capturados e alguns dias neste intervalo. Os números do eixo vertical, ou **eixo-*y***, indicam a **frequência**: o número de observações que tiveram um resultado particular. Por exemplo, houve apenas um dia no qual 4 capturas foram registradas; dois

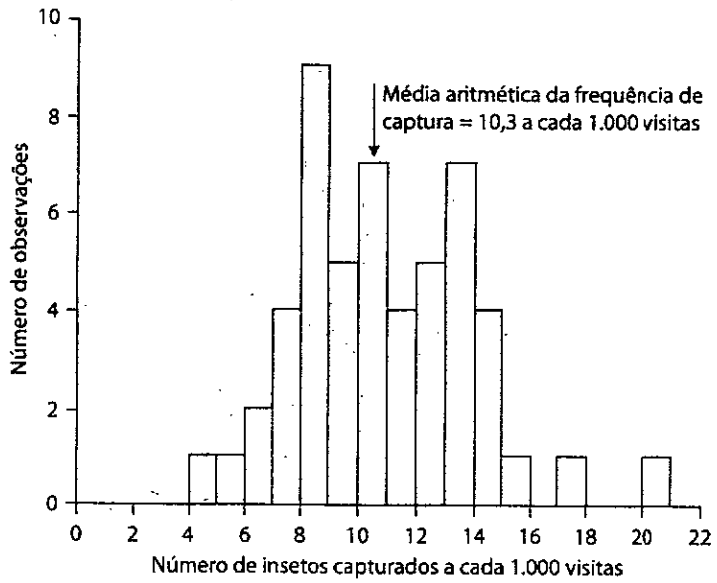


Figura 1.1 Histograma da frequência de capturas. Em cada semana, por um ano inteiro, um investigador poderia observar a visita de 1.000 insetos a uma planta-jarro carnívora e contar quantas vezes a planta capturou um inseto (*ver* Tabela 1.1). Os dados coletados de todas as 52 observações (uma por semana) podem ser plotados na forma de um histograma. O número de capturas observado varia de um mínimo de 4 em uma semana (coluna na extrema esquerda) a um máximo de 20 em outra (coluna da extrema direita); a média aritmética de todas as 52 observações é 10,3 capturas a cada 1.000 visitas.

[6] Apesar de na verdade não termos conduzido esse experimento, tais dados foram coletados por Newell & Nastase (1998), que filmaram a visita de insetos às plantas-jarro e registraram 27 capturas em 3.308 visitas, das quais 74 tiveram um destino desconhecido. Para as análises deste capítulo, simulamos dados usando um gerador de números aleatórios em uma planilha de computador. Para essa simulação, estabelecemos a probabilidade "verdadeira" de captura em 0,01, o tamanho amostral em 1.000 visitas e o número de amostras em 52. Usamos, então, os dados resultantes para gerar o histograma da Figura 1.1. Se você conduzir o mesmo experimento de simulação no seu computador, obteria resultados similares, mas haveria alguma variabilidade, dependendo de como seu computador gera números aleatórios. Discutiremos experimentos de simulação com mais detalhes nos Capítulos 3 e 5.

dias com 6 capturas registradas; e cinco dias com 9 capturas. Uma maneira de estimar a probabilidade de captura é tirando a média aritmética de todas as amostras. Em outras palavras, poderíamos calcular o número médio de capturas em cada uma das 52 amostras de 1.000 visitas. Neste exemplo, o número médio de capturas dentre essas visitas é 10,3. Assim, a probabilidade de ser capturado é de 10,3/1.000 = 0,0103, ou somente um pouco mais de um em 100. Chamamos essa média aritmética de *valor esperado da probabilidade*, ou **expectativa**, e a denotamos como $E(P)$. Distribuições como a mostrada na Figura 1.1 são frequentemente utilizadas para descrever os resultados de experimentos em probabilidade. Retornaremos a eles, em maiores detalhes, no Capítulo 3.

PROBLEMAS NA DEFINIÇÃO DE PROBABILIDADE

A maioria dos livros-texto de estatística define muito rapidamente probabilidade apenas como fizemos aqui: a frequência (esperada) na qual eventos ocorrem. Os exemplos padrão dos livros-texto são aqueles de que o lançamento de uma moeda honesta tem uma probabilidade de 0,50 (igual a uma chance de 50%) de sair cara,[7] ou o de que cada face de um dado honesto de seis lados tem a chance de 1 em 6 de cair para cima. Porém, vamos considerar esses exemplos com mais cuidado. Por que aceitamos esses valores como as probabilidades corretas?

Aceitamos essas respostas por causa da nossa visão de como moedas são lançadas e de como os dados são rolados. Por exemplo, se equilibrarmos uma moeda no topo de uma mesa com a "cara" virada para cima e então tombarmos suavemente o objeto para frente, todos concordaríamos que a probabilidade de obter cara não mais será de 0,50. Em vez disso, a probabilidade de 0,5 se aplica somente a uma moeda que é lançada no ar.[8] Mesmo assim, deveríamos reconhecer que a probabilidade de 0,50 representa realmente uma estimativa baseada em uma quantidade mínima de dados.

Por exemplo, suponha que temos uma vasta gama de microssensores de alta tecnologia aderida à superfície da moeda, nos músculos de nossa mão e nas

[7] Em um curioso lance do destino, uma das novas moedas de Euro pode não ser honesta. A moeda do Euro, introduzida em 2002 por 12 estados europeus, tem um mapa da Europa no lado da "coroa", porém cada país tem seu próprio desenho no lado da "cara". Os estatísticos poloneses Tomasz Gliszczynski e Waclaw Zawadowski lançaram (na verdade rodopiaram) o Euro belga 250 vezes e obtiveram cara em 56% das vezes (140 caras). Eles atribuíram este resultado a uma imagem pesada embutida no lado "cara", mas Howard Grubb, um estatístico da Universidade de Reading, apontou que de 250 observações, 56% "não é significativamente diferente" de 50%. Quem está certo? (como reportado em *New Scientist*: http://www.newscientist.com/, 2 de Janeiro de 2002). Retornaremos a este exemplo no Capítulo 11.

[8] Apostadores e trapaceiros tentam contrariar este modelo usando moedas ou dados viciados, ou arremessando moedas ou dados cuidadosamente, de forma a favorecer uma das faces, mas dando a aparência de aleatoriedade. Cassinos trabalham em vigília para garantir que um modelo aleatório seja executado; dados devem ser chacoalhados com vigor, as roletas giradas bem rápido, e as cartas do baralho de vinte um (*blackjack*) são frequentemente embaralhas. Não é necessário que o cassino trapaceie. As tabelas de prêmios são calculadas para produzir um belo lucro, e tudo que o cassino tem a faze: é garantir que os clientes não possam influenciar ou prever o resultado de cada jogo.

paredes da sala em que estamos. Esses sensores detectariam e quantificariam a exata quantidade de torque no nosso lançamento, a temperatura e turbulência do ar na sala, as microirregularidades na superfície da moeda e os microturbilhões na turbulência do ar que são gerados conforme a moeda gira. Se todos esses dados fossem instantaneamente passados para um computador, poderíamos desenvolver um modelo muito complexo capaz de descrever a trajetória da moeda. Com tal informação, poderíamos prever, com muito mais acurácia, qual face da moeda irá cair para cima. De fato, se tivéssemos uma quantidade infinita de dados disponíveis, talvez não houvesse nenhuma incerteza no resultado do lançamento da moeda.[9] No fim das contas, as observações que usamos para estimar a probabilidade de um evento podem não ser tão similares. Cada observação representa um conjunto de condições único; uma cadeia de causa e efeito que determina inteiramente se o resultado será cara ou coroa. Se pudéssemos duplicar com perfeição esse conjunto de condições, não haveria nenhuma incerteza no resultado do evento e absolutamente nenhuma necessidade de usar probabilidade ou estatística![10]

Werner Heisenberg

Erwin Schrödinger

[9] Esta última declaração é bastante debatida. Ela assume que medimos todas as variáveis certas e que as interações entre aquelas são simples o suficiente para podermos mapear todas as contingências ou as expressar em uma relação matemática. A maioria dos cientistas acredita que modelos mais complexos são mais acurados que os mais simples, mas que pode não ser possível eliminar toda a incerteza. Note, também, que modelos complexos não nos ajudam se não podemos medir as variáveis que eles incluem; é improvável que a tecnologia necessária para monitorar nosso arremesso da moeda exista em breve. Por fim, existe o problema mais súbito de que o próprio ato de medir as coisas na natureza pode alterar o processo que estamos tentando estudar. Por exemplo, suponha que desejamos quantificar a abundância relativa de espécies de peixes próximo ao fundo do oceano, onde a luz solar nunca penetra. Podemos usar câmeras subaquáticas para fotografar peixes e então contar o número de peixes em cada fotografia. A luz da câmera, porém, atrai algumas espécies e repele outras. O que, exatamente, medimos? Na física, o Princípio da Incerteza, de Heisenberg (nome atribuído após a morte do físico alemão Werner Heisenberg), diz que é impossível medir, de maneira simultânea, a posição e o momento de um elétron: quanto mais acuradamente você mede uma dessas quantidades, menos você consegue medir a outra. Se o próprio ato de observar fundamentalmente causa mudanças em processos na natureza, então a cadeia de causa e efeito é quebrada (ou no mínimo distorcida). Tal conceito foi formulado por Erwin Schrödinger (1887-1961), que colocou (teoricamente) um gato e um frasco de cianeto em uma caixa de quantum. Dentro da caixa, o gato pode estar simultaneamente vivo ou morto, porém uma vez que a caixa seja aberta e o gato observado, ele estará ou vivo ou morto eternamente.

[10] Muitos cientistas poderiam adotar essa falta de incerteza. Alguns de nossos colegas biólogos moleculares estariam muito felizes em um mundo no qual a estatística fosse desnecessária. Quando há incerteza no resultado de seus experimentos, eles frequentemente assumem que foi devido a uma técnica mal empregada e que, eliminando os erros de medidas e a contaminação, obterão dados limpos, repetíveis e corretos. Os ecólogos e biólogos de campo em geral são mais veementes acerca da variação em seus sistemas. Não só porque os sistemas ecológicos são inerentemente mais residuais que os moleculares; em vez disso, dados ecológicos podem ser mais variáveis que moleculares, porque, com frequência, são coletados em diferentes escalas espaciais e temporais.

Retornando às plantas-jarro, é óbvio que nossas estimativas da probabilidade de captura irão depender muito dos detalhes de cada observação. As chances de captura diferem entre as plantas grandes e pequenas, entre as presas formigas e as presas voadoras, e entre plantas ao sol e à sombra. E, para cada um desses grupos, poderíamos fazer subdivisões adicionais baseadas em detalhes ainda mais refinados acerca das condições da visita. Nossas estimativas de probabilidade irão depender de quão ampla ou restritamente limitaremos os tipos de observações que serão consideradas.

Resumindo, quando dizemos que eventos são aleatórios, estocásticos, probabilísticos ou devido à chance, o que realmente acontece é que os resultados são determinados, em parte, por um conjunto complexo de processos que somos incapazes ou relutantes em medir e o tratamos como aleatório. A força de outros processos que podemos medir, manipular e modelar representam forças determinísticas ou mecanicistas. Podemos pensar nos padrões em nossos dados como sendo determinados por um misto de forças determinísticas e aleatórias. Como observadores, impomos uma distinção entre determinístico e aleatório. Isso reflete em nosso modelo conceitual implícito acerca das forças em operação, e na disponibilidade de dados e medidas que são relevantes para essas forças.[11]

A MATEMÁTICA DA PROBABILIDADE

Nesta sessão, apresentamos um breve tratamento matemático sobre como calcular probabilidades. Apesar de o conteúdo parecer tangencial à indagação (*os dados são "significativos"*?), a correta interpretação de resultados estatísticos se articula sobre essas operações.

Definindo o espaço amostral

Como ilustrado no exemplo da captura de presas pelas plantas carnívoras, a probabilidade P de ser capturado (o resultado) enquanto visita uma planta (o evento) é definida simplesmente como o número de capturas dividido pelo de visitas (o

[11] Se rejeitarmos a noção de qualquer aleatoriedade na natureza, rapidamente nos movemos para o domínio filosófico do misticismo: "Qual é, então, a Lei do Carma? A Lei, sem exceção, que rege todo o universo, do invisível e imponderável átomo aos sóis; ... e esta lei diz que *toda causa produz um efeito*, sem qualquer possibilidade de atrasar ou anular o efeito, uma vez que a causa começa a operar. A lei da causalidade é suprema em qualquer lugar. Eis a Lei do Carma; carma é a ligação inevitável entre causa e efeito." (Arnould 1895).

A ciência ocidental não progride bem incorporando esse tipo de complexidade que tudo – engloba. É verdade que muitas explicações científicas para fenômenos naturais são complicadas. Contudo, essas explicações são obtidas, em princípio, evitando toda complexidade e estabelecendo uma **hipótese nula** que tenta explicar padrões nos dados da forma mais simples possível, o que significa, com frequência, atribuir inicialmente a variação nos dados à aleatoriedade (ou erro de medição). Se essa hipótese nula simples pode ser rejeitada, podemos passar a nos entreter com hipóteses mais complexas. *Ver* Capítulo 4 para mais detalhes sobre hipóteses nulas e teste de hipóteses. *Ver, também* Beltrami (1999) para uma discussão sobre aleatoriedade.

número de observações) (Equação 1.1). Examinaremos em detalhes cada um desses termos: resultado, evento, observação e probabilidade.

Primeiro, precisamos definir nosso universo de eventos possíveis ou o espaço amostral de interesse. Em nosso primeiro exemplo, um inseto poderia visitar com sucesso uma planta e sair, ou poderia ser capturado. Esses dois possíveis resultados formam o espaço amostral (ou conjunto), que chamamos de *Visita*:

$$Visita = \{(Captura), (Fuga)\}$$

Usamos chaves {} para simbolizar o conjunto e parênteses () para simbolizar os eventos de um conjunto. Os objetos dentro dos parênteses são resultados do evento. Como há somente dois resultados possíveis para uma visita, se a probabilidade do resultado *Captura* é 1 em 1.000, ou 0,001, então a probabilidade de fuga é de 999 em 1.000, ou 0,999. Este exemplo simples pode ser generalizado ao **Primeiro Axioma da Probabilidade:**

Axioma 1: A soma de todas as probabilidades dos resultados em um único espaço amostral = 1,0.

Podemos escrever esse axioma como

$$\sum_{i=1}^{n} P(A_i) = 1{,}0$$

que lemos como "a soma das probabilidades de todos os resultados A_i é igual a 1,0." O "W de lado" (letra grega maiúscula *sigma*) é o sinal de somatório, a notação taquigráfica para somar os elementos a seguir. O *i* subscrito indica um elemento em particular e significa que iremos somar os elementos de $i = 1$ até n, onde n é o número total de elementos no conjunto. Em um espaço amostral propriamente definido, dizemos que os eventos são **mutuamente exclusivos** (um indivíduo ou é capturado ou escapa), e que os resultados são **completos** (*Captura* ou *Fuga* são os únicos resultados possíveis do evento). Se os eventos não são mutuamente exclusivos ou completos, as probabilidades dos eventos no espaço amostral não somarão 1,0.

Em muitos casos, haverá mais que somente dois resultados possíveis para um evento. Por exemplo, em um estudo hipotético da reprodução de um besouro girinídeo de manchas na cor laranja, estabelecemos que cada um desses animais sempre produz exatamente 2 proles, com 2 a 4 descendentes em cada uma. O sucesso reprodutivo no tempo de vida de um girinídeo de manchas alaranjadas pode ser descrito como um resultado (a,b), onde a representa o número de descendentes da primeira prole e b o número de descendentes da segunda. O espaço amostral *Aptidão* consiste em todos os resultados reprodutivos que um indivíduo pode obter durante sua vida:

$$Aptidão = \{(2,2), (2,3), (2,4), (3,2), (3,3), (3,4), (4,2), (4,3), (4,4)\}$$

Como cada girinídeo pode gerar apenas 2, 3 ou 4 descendentes por prole, esses 9 pares de números inteiros são os únicos resultados possíveis. Na ausência de qualquer outra informação, nós inicialmente partimos da premissa simplificada de que as probabilidades de cada um desses resultados reprodutivos são iguais. Usamos a definição de probabilidade da Equação 1.1 (número de resultados/número de observações) para determinar este valor. Visto que existem 9 resultados possíveis neste conjunto, $P(2,2) = P(2,3) = ... = P(4,4) = 1/9$. Note também que as probabilidades obedecem ao Axioma 1: a soma das probabilidades de todos os resultados = 1/9 + 1/9 + 1/9 + 1/9 + 1/9 + 1/9 + 1/9 + 1/9 + 1/9 = 1,0.

Eventos complexos e compartilhados: combinando probabilidades simples

Uma vez que as probabilidades de eventos simples são conhecidas ou foram estimadas, podemos usá-las para medir probabilidades de eventos mais complicados. Os **eventos complexos** são compostos por eventos simples do espaço amostral. Os **eventos compartilhados** são ocorrências múltiplas e simultâneas de eventos simples no espaço amostral. As probabilidades de eventos complexos e compartilhados podem ser decompostos nas somas ou produtos das probabilidades dos eventos simples. Contudo, pode ser difícil decidir quando as probabilidades devem ser adicionadas ou ser multiplicadas. A resposta pode ser obtida determinando se o novo evento pode ser atingido por uma das várias rotas diferentes (evento complexo), ou se ele requer a ocorrência simultânea de dois ou mais eventos simples (evento compartilhado).

Se o novo evento pode ocorrer por diferentes rotas, ele é complexo e pode ser representado como uma **declaração *ou***: Evento A *ou* Evento B *ou* Evento C. Desta forma, se diz que eventos complexos representam a **união** de eventos simples. As probabilidades de eventos complexos são determinadas somando as de eventos simples.

Em contraste, se novos eventos requerem a ocorrência simultânea de vários eventos simples, ele é um evento compartilhado e pode ser representado como uma **declaração *e***: Evento A *e* Evento B *e* Evento C. Eventos compartilhados, portanto, representam a **interseção** de eventos simples. As probabilidades de eventos compartilhados são determinadas pela multiplicação das probabilidades.

EVENTOS COMPLEXOS: SOMANDO PROBABILIDADES O exemplo do girinídeo pode ser utilizado para ilustrar o cálculo da probabilidade de eventos complexos. Suponha que queremos medir o ganho da vida reprodutiva de um besouro girinídeo. Poderíamos contar o número de descendentes que um girinídeo produz em sua vida. Esse número é a soma dos descendentes das duas proles, o que resulta em um número inteiro entre 4 e 8. Como poderíamos determinar a probabilidade de um girinídeo de manchas alaranjadas produzir 6 descendentes? Primeiro, note que o evento que produz *6 descendentes* pode ocorrer de três formas:

$$6\ descendentes = \{(2,4), (3,3), (4,2)\}$$

e esse evento complexo é, por si só, um conjunto.

Podemos ilustrar esses dois conjuntos com um **diagrama de Venn.**[12] A Figura 1.2 ilustra o diagrama de Venn para nossos conjuntos de *Aptidão* e de *6 descendentes*. Graficamente, você pode ver que *6 descendentes* é um **subconjunto próprio** do conjunto maior *Aptidão* (ou seja, todos os elementos do primeiro são elementos do último). Indicamos que um conjunto é um subconjunto do outro conjunto com o símbolo $\subset$, e neste caso podemos escrever

$$6\ descendentes \subset Aptidão$$

Três dos 9 resultados possíveis de *Aptidão* produzem 6 descendentes, então podemos estimar a probabilidade de ter 6 descendentes como sendo 1/9 + 1/9 + 1/9 = 3/9 (relembrando nossa premissa de que cada resultado é equiprovável). Tal resultado é generalizado no **Segundo Axioma de Probabilidade**:

Axioma 2: A probabilidade de um evento complexo é igual à soma das probabilidades dos resultados que compõem aquele evento.

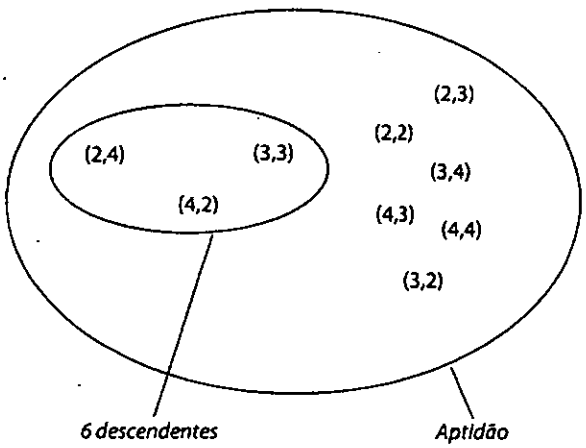


Figura 1.2 Diagrama de Venn ilustrando o conceito de conjuntos. Cada par de números representa o número de descendentes produzidos em duas proles consecutivas de uma espécie hipotética do besouro giriníndeo. Assumimos que esse besouro produz exatamente 2, 3 ou 4 descendentes a cada vez que se reproduz, então estes são os únicos números inteiros que aparecem no diagrama. Um conjunto em um diagrama de Venn é um anel que compreende certos elementos. O conjunto *Aptidão* contém todos os resultados reprodutivos possíveis para duas proles consecutivas. Dentro do conjunto *Aptidão*, há um conjunto menor *6 descendentes*, constituído pelas proles que produzem exatamente um total de 6 descendentes (i.e., cada par de números inteiros somam 6). Dizemos que *6 descendentes* é um subconjunto próprio de *Aptidão*, pois os elementos do primeiro estão completamente contidos no segundo.

John Venn

[12] John Venn (1834-1923) estudou em Gonville and Caius College, Cambridge University, Inglaterra, onde se graduou em 1857. Ele foi ordenado padre dois anos depois e retornou a Cambridge em 1562, como professor de Ciência Moral. Ele também estudou e lecionou lógica e teoria de probabilidades. Ele é mais conhecido pelos diagramas que representam conjuntos, suas uniões e interseções, e agora esses diagramas levam seu nome. Venn também é lembrado por construir uma máquina para arremessar bolas de *cricket*.

Você pode pensar em um evento complexo, como uma declaração *ou* em um programa de computador: se os eventos simples são A, B e C, o evento complexo é (A ou B ou C), pois qualquer um desses resultados irá representar um evento complexo. Assim

$$P(A \text{ ou } B \text{ ou } C) = P(A) + P(B) + P(C) \qquad (1.2)$$

Como outro exemplo simples, considere a probabilidade de retirar apenas uma carta de um baralho de 52 cartas e obter um Ás. Sabemos que há 4 azes no baralho, e a probabilidade do evento complexo de tirar qualquer um dos 4 azes é de:

$$P(Ás) = 1/52 + 1/52 + 1/52 + 1/52 = 4/52 = 1/13$$

EVENTOS COMPARTILHADOS: MULTIPLICANDO PROBABILIDADES No exemplo do besouro girinídeo, calculamos a probabilidade de um girinídeo produzir exatamente 6 descendentes. Este evento complexo poderia ocorrer por qualquer 1 de 3 diferentes pares de proles {(2,4), (3,3), (4,2)} e determinamos a probabilidade somando as probabilidades simples.

Agora, passemos para o cálculo da probabilidade de um evento compartilhado – a ocorrência simultânea de dois eventos simples. Assumimos que o número de descendentes produzidos na segunda prole é independente do número de descendentes produzidos na primeira. A **independência** é uma premissa simplificada que é crítica em muitas análises estatísticas. Quando dizemos que dois eventos são independentes um do outro, significa que o resultado de um não é afetado pelo do outro. Se dois eventos são independentes um do outro, a probabilidade de que ambos ocorram (evento compartilhado) é igual ao produto de suas probabilidades individuais.

$$P(A \cap B) = P(A) \times P(B) \text{ (se } A \text{ e } B \text{ são independentes)} \qquad (1.3)$$

O símbolo $\cap$ indica a **interseção** de dois eventos independentes – ou seja, ambos ocorrendo simultaneamente. Por exemplo, no exemplo do tamanho da prole, suponha que um indivíduo pode produzir 2, 3, 4 descendentes na primeira prole e que a chance de cada um desses eventos é de 1/3. Se as mesmas regras se aplicam para a produção da segunda prole, a probabilidade de obter um par de proles (2,4) é igual a $1/3 \times 1/3 = 1/9$. Note que este é o mesmo número obtido ao tratar cada um dos diferentes pares de prole como independentes, eventos equiprováveis.

Cálculo de probabilidades: paineirinhas* e lagartas

Trata-se de um exemplo simples que incorpora tanto eventos complexos quanto compartilhados. Imagine um conjunto de afloramentos rochosos no qual pode-

* N. de T. No original, *milkweeds* é o nome dado a herbáceas da família Apocynaceae, geralmente pertencentes ao gênero *Asclepias*. Muitas espécies são invasoras e foram introduzidas no Brasil. Podem receber muitos nomes populares, como cega-olho, erva-leiteira, mata-rato, flor-borboleta, paina de seda e paineirinha.

mos encontrar populações de paineirinha, bem como lagartas herbívoras. Duas variedades de populações de paineirinha estão presentes: aquelas que evoluíram um composto secundário que as tornam resistentes (R) aos herbívoros e aquelas que não possuem. Suponha que você faz o censo de um número de populações de paineirinhas e determina que $P(R) = 0{,}20$. Em outras palavras, 20% das populações de paineirinha são resistentes aos herbívoros. Os outros 80% das populações remanescentes representam o complemento desse conjunto. O complemento inclui todos os outros elementos do conjunto, que podemos escrever brevemente como *não R*. Assim

$$P(R) = 0{,}20 \qquad P(\text{não } R) = 1 - P(R) = 0{,}80$$

Similarmente, suponha que a probabilidade de que a lagarta (L) ocorra em uma mancha é de 0,7:

$$P(C) = 0{,}7 \qquad P(\text{não} C) = 1 - P(C) = 0{,}30$$

A seguir, especificamos as regras ecológicas que determinam a interação entre lagartas e paineirinhas e então usamos a teoria de probabilidades para determinar as chances de encontrar lagartas, paineirinhas ou ambas presentes nas manchas. As regras são simples. Primeiramente, todas as paineirinhas e lagartas podem dispersar e chegar a todas as manchas de afloramentos rochosos. As populações de paineirinhas sempre persistem quando as lagartas não estão presentes, mas, quando presente, somente as populações de paineirinhas resistentes persistem. Como antes, assumimos que as paineirinhas e as lagartas colonizam inicialmente as manchas de maneira independente umas das outras.[13]

Primeiro, vamos considerar as diferentes combinações de populações resistentes e não resistentes ocorrendo com e sem herbívoros. São dois eventos simultâneos, então iremos multiplicar probabilidades para gerar os 4 eventos compartilhados possíveis (Tabela 1.2). Existem algumas coisas importantes a se notar na Tabela 1.2. Primeiro, a soma das probabilidades resultantes dos eventos compartilhados (0,24 + 0,56 + 0,06 + 0,14) = 1,0, e estes 4 eventos compartilhados juntos formam um conjunto próprio. Segundo, podemos somar algumas dessas probabilidades para definir novos eventos complexos e também para recuperar alguma das probabilidades simples subjacentes. Por exemplo, qual a probabilidade de que populações de paineirinhas sejam resistentes? Isto pode ser estimado como a pro-

[13] Muita biologia interessante surge quando a premissa de independência é violada. Por exemplo, muitas espécies de borboletas e mariposas adultas são bastante seletivas e buscam manchas com plantas hospedeiras apropriadas para depositar seus ovos. Consequentemente, a ocorrência de lagartas pode não ser independente da planta hospedeira. Em outro exemplo, a presença de herbívoros aumenta a pressão seletiva para a evolução de resistência nos hospedeiros. Além disso, muitas espécies de plantas possuem os tão falados defensores químicos facultativos que são "ligados" somente quando os herbívoros aparecem. Além disso, a ocorrência de populações resistentes pode não ser independente da presença de herbívoros. Mais à frente, neste capítulo, aprenderemos alguns métodos que incorporam probabilidades não independentes em nossos cálculos de eventos complexos.

TABELA 1.2 Cálculos de probabilidade para eventos compartilhados

Evento compartilhado	Cálculo da probabilidade	Resultado: Paineirinha presente?	Resultado: Lagarta presente?
População suscetível sem lagarta	$[1 - P(R) \times 1 - P(L)] = (1,0 - 0,2) \times (1,0 - 0,7) = 0,24$	Sim	Não
População suscetível com lagarta	$[1 - P(R) \times P(L)] = (1,0 - 0,2) \times (0,7) = 0,56$	Não	Sim
População resistente sem lagarta	$[P(R) \times 1 - P(L)] = (0,2) \times (1,0 - 0,7) = 0,06$	Sim	Não
População resistente com lagarta	$[P(R) \times P(L)] = (0,2) \times (0,7) = 0,14$	Sim	Sim

A ocorrência conjunta de eventos independentes é um evento compartilhado e sua probabilidade pode ser calculada como o produto das probabilidades dos eventos individuais. Neste exemplo hipotético, populações de paineirinhas que são suscetíveis às lagartas herbívoras ocorrem com probabilidade $P(R)$, e populações de paineirinhas que são resistentes às lagartas ocorrem com probabilidade $1 - P(R)$. A probabilidade de que uma mancha de paineirinhas seja colonizada por lagartas é $P(L)$ e a de que uma mancha de paineirinhas não seja colonizada por lagartas é de $1 - P(L)$. Os eventos simples são a ocorrência de lagartas (L) e de populações resistentes de paineirinhas (R). A primeira coluna lista os quatro eventos complexos, definidos por populações resistentes ou suscetíveis de paineirinhas e pela presença ou ausência de lagartas. A segunda coluna ilustra o cálculo da probabilidade do evento complexo. A terceira e quarta colunas indicam o resultado ecológico. Note que as populações de paineirinhas são extintas se forem suscetíveis e colonizadas por lagartas. A probabilidade de que evento ocorra é de 0,56, deste modo esperamos esse resultado 56% das vezes. O resultado mais improvável é uma população de paineirinhas resistente que não contenha lagartas ($P = 0,06$). Note também que os quatro eventos compartilhados formam um conjunto próprio, portanto suas probabilidades somam 1,0 (0,24 + 0,56 + 0,06 + 0,14 = 1,0).

babilidade de encontrar populações resistentes com lagartas ($P = 0,14$) mais a probabilidade de encontrar populações resistentes sem lagartas ($P = 0,06$). A soma (0,20) de fato condiz com a probabilidade original de resistência [$P(R) = 0,20$]. A independência dos dois eventos garante que podemos recuperar os valores originais desta forma.

Todavia, também aprendemos algo novo com esse exercício. As paineirinhas irão desaparecer se uma população suscetível for colonizada por lagartas, e isso pode ocorrer a uma probabilidade de 0,56. O complemento deste evento, (1 - 0,56) = 0,44, é a probabilidade de que um local irá conter uma população de paineirinhas. Equivalentemente, a probabilidade de persistência das paineirinhas pode ser calculada como P(paineirinha presente) = 0,24 + 0,06 + 0,14 = 0,44, somando as probabilidades das diferentes combinações que resultam na presença de paineirinhas. Assim, apesar da probabilidade de resistência ser de apenas 0,20, esperamos encontrar populações de paineirinhas em 44% das manchas amostradas, porque nem todas as populações suscetíveis são atingidas por lagartas. Novamente, enfatizamos que tais cálculos são corretos apenas se os eventos iniciais de colonização forem independentes um do outro.

Eventos complexos e compartilhados: regras combinatórias de conjuntos

Muitos eventos não são independentes uns dos outros. Não obstante, precisamos de métodos para levar em conta a falta de independência. Retornando ao exemplo do besouro girinídeo, como seria se o número de descendentes da segunda prole de alguma forma fosse relacionado ao produzido na primeira prole? Isso pode ocorrer porque organismos possuem uma quantidade limitada de energia disponível para produzir descendentes, de forma que a energia investida na primeira prole não está disponível para investimento na segunda prole. Isso afetaria nossa estimativa da probabilidade de produzir 6 descendentes? Antes que possamos responder, precisamos de mais ferramentas em nossa caixa de ferramentas probabilísticas. Essas ferramentas nos informam como combinar eventos ou conjuntos e nos permite calcular as probabilidades de eventos combinatórios.

Suponha que em nosso espaço amostral existem dois eventos identificáveis, e que cada um consiste de um grupo resultados. Por exemplo, no espaço amostral *Aptidão*, poderíamos descrever um evento consistindo em um besouro que produz exatamente 2 descendentes em sua primeira prole. Iremos chamar esse evento de *primeira prole 2*, e abreviar como *F*. O segundo evento é um besouro que produz exatamente 4 descendentes na segunda prole. Chamaremos de *segunda prole 4* e abreviaremos como *S*:

$$\textit{Aptidão} = \{(2,2), (2,3), (2,4), (3,2), (3,3), (3,4), (4,2), (4,3), (4,4)\}$$
$$F = \{(2,2), (2,3), (2,4)\}$$
$$S = \{(2,4), (3,4), (4,4)\}$$

Podemos construir dois novos conjuntos a partir de *F* e *S*. O primeiro é o novo conjunto de resultados, que compreende todos os resultados que fazem parte de *F* ou de *S*. Indicamos esse novo conjunto usando a notação $F \cup S$ e o chamamos de **união** dos dois conjuntos.

$$F \cup S = \{(2,2), (2,3), (2,4), (3,4), (4,4)\}$$

Note que o resultado (2,4) ocorre em ambos, *F* e *S*, mas é contado apenas uma vez em $F \cup S$. Note, também, que a união de *F* e *S* é um conjunto que contém mais elementos que *F* e *S* possuem individualmente, pois esses conjuntos são "adicionados" para criar a união.

O segundo novo conjunto é composto pelos resultados que estão em *F* e *S*. Indicamos esse conjunto com a notação $F \cap S$ e o chamamos de **interseção** de dois conjuntos.

$$F \cap S = \{(2,4)\}$$

Note que a interseção entre *F* e *S* é um conjunto que contém menos elementos, os quais estão contidos apenas em *F* ou *S*, pois agora estamos considerando apenas os elementos que são comuns a ambos os conjuntos. O diagrama de Venn, na Figura 1.3, ilustra estas operações de união e interseção.

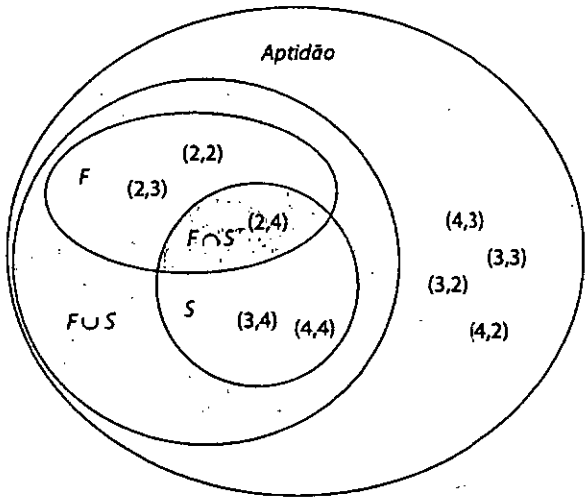


Figura 1.3 Diagrama de Venn ilustrando uniões e interseções de conjuntos. Cada anel representa um conjunto diferente do número de descendentes produzidos por um par de proles de uma espécie hipotética do besouro girinídeo, como na Figura 1.2. O anel maior representa o conjunto *Aptidão*, que compreende todos os resultados reprodutivos possíveis do besouro. O anel menor F é o conjunto de todos os pares de proles com exatamente dois descendentes na primeira prole. O anel menor S é o conjunto de todos os pares de proles com exatamente 4 descendentes na segunda prole. A área na qual os anéis se sobrepõem representa a interseção entre F e S ($F \cap S$) e contém os elementos comuns a ambos. O anel que compreende F e S representa a união entre F e S ($F \cup S$) e contém todos os elementos de ambos os conjuntos. Note que a união de F e S não duplica o elemento em comum (2,4). Em outras palavras, a união de dois conjuntos é a soma dos elementos de ambos, menos seus elementos em comum. Assim, $(F \cup S) = (F + S) - (F \cap S)$.

Podemos construir um terceiro conjunto útil considerando o F^c, chamado de complemento de F, que é o conjunto dos objetos remanescentes no espaço amostral (*Aptidão*, nesse exemplo) que não estão em F:

$$F^c = \{(3,2), (3,3), (3,4), (4,2), (4,3), (4,4)\}$$

A partir dos Axiomas 1 e 2, você deveria ver que:

$$P(F) + P(F^c) = 1{,}0$$

Em outras palavras, como F e F^c coletivamente incluem todos os resultados possíveis (por definição, eles compreendem o conjunto inteiro), a soma das probabilidades associadas a eles deve resultar em 1,0.

Por fim, precisamos apresentar o **conjunto vazio**. Ele não contém elementos e é escrito como $\{\varnothing\}$. Por que esse conjunto é importante? Considere o conjunto que consiste da interseção entre F e F^c. Como não possuem nenhum elemento em comum, se não tivéssemos um conjunto vazio, então $F \cap F^c$ seria indefinido. Um

conjunto vazio permite que os conjuntos sejam fechados[14] sob as três operações permitidas: união, interseção e complemento.

CALCULANDO A PROBABILIDADE DE EVENTOS COMBINATÓRIOS o Axioma 2 declara que a probabilidade de um evento complexo é igual à soma das probabilidades dos resultados que compreendem um evento. Esta pode ser uma operação simples; portanto, para calcular a probabilidade, digamos, $F \cup S$ como $P(F) + P(S)$. Como F e S possuem 3 resultados cada, com probabilidade de 1/9, a soma simples resulta em 6/9. Contudo, existem apenas 5 elementos em $F \cup S$, então a probabilidade deve ser somente de 5/9. Por que o axioma não funciona? A Figura 1.3 mostra que esses dois conjuntos possuem o resultado (2,4) em comum. Não é uma coincidência que este total seja igual à interseção entre F e S. O cálculo simples da união desses dois conjuntos resultou na contagem dupla do elemento em comum. Em geral, precisamos evitar a contagem dupla quando calculamos a probabilidade em uma união.[15]

Portanto, calculamos a probabilidade de uma união de quaisquer 2 conjuntos ou eventos A e B como:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \qquad (1.4)$$

[14] **Fechamento** é uma propriedade matemática. Uma coleção de objetos G (um **grupo**, tecnicamente) é fechada sob uma operação $\oplus$ (como uma união, uma interseção ou um complemento) se, para todos os elementos A e B de G, $A \oplus B$ também seja um elemento de G.

[15] Determinar se um evento foi contado duas vezes ou não pode ser complicado. Por exemplo, em genética básica de populações, a familiar equação de Hardy-Weinberg fornece as frequências de diferentes genótipos em uma população com acasalamentos aleatórios. Para um único gene, sistema de dois alelos, a equação de Hardy-Weinberg prediz a frequência de cada genótipo (*RR*, *Rr* e *rr*) em uma população na qual p é a frequência inicial do alelo R e q é a frequência inicial do alelo r. Como os genótipos R e r formam um conjunto fechado, $p + q = 1{,}0$. Cada parente contribui com um único alelo em seus gametas, de forma que a formação de descendentes é um evento compartilhado, com dois gametas se combinando na fertilização e contribuindo com seus alelos para determinar o genótipo dos descendentes. A equação de Hardy-Weinberg prediz a probabilidade (ou frequência) das diferentes combinações de genótipos dos descendentes em uma população com acasalamentos aleatórios como

$$P(RR) = p^2$$

$$P(Rr) = 2pq$$

$$P(rr) = q^2$$

A produção de um genótipo *RR* requer que ambos os gametas contenham o alelo R. Como a frequência do alelo R na população é p, a probabilidade desse evento complexo = $p \times p = p^2$. Então, por que o genótipo heterozigoto (*Rr*) ocorre a uma frequência de $2pq$ e não de apenas pq? Não é um caso de contagem dupla? A resposta é "não", pois há duas formas de criar um indivíduo heterozigoto: podemos combinar o gameta R de um macho com o r de uma fêmea, ou combinar o r de um macho com o R de uma fêmea. Esses dois eventos são distintos, mesmo que o zigoto resultante tenha o mesmo genótipo (*Rr*) em ambos os casos. Portanto, a probabilidade da formação de um heterozigoto é um evento complexo, cujos elementos são dois eventos compartilhados:

$$P(Rr) = P(\text{macho } R \text{ e fêmea } r) \text{ ou } P(\text{macho } r \text{ e fêmea } R)$$

$$P(Rr) = pq + qp = 2pq$$

A probabilidade de uma interseção é a mesma de ambos os eventos ocorrerem. No exemplo mostrado na Figura 1.3, $F \cap S = \{(2,4)\}$, cuja probabilidade é 1/9, o produto das de F e de S. ((1/3) × (1/3) = 1/9). Assim, a subtração de 1/9 menos 6/9 resulta na probabilidade desejada de 5/9 para $P(F \cup S)$.

Se não houver sobreposição entre A e B, então eles são **eventos mutuamente exclusivos**, significando a falta de resultados em comum. A interseção de dois eventos mutuamente exclusivos ($A \cap B$) é, por consequência, o conjunto vazio $\{\varnothing\}$. Como este não possui nenhum resultado e tem probabilidade = 0, podemos simplesmente somar as probabilidades de dois eventos mutuamente exclusivos para obter a de sua união:

$$P(A \cup B) = P(A) + P(B) \quad (\text{se } A \cap B = \{\varnothing\}) \tag{1.5}$$

Agora estamos prontos para retornar à questão de estimar a probabilidade de um besouro girinídeo de manchas cor de laranja produzir 6 descendentes, caso o número de descendentes produzidos na segunda prole dependa do número de descendentes da primeira. Relembre que o evento complexo *6 descendentes* consiste de três resultados (2,4), (3,4), (4,2); sua probabilidade é $P(6\ descendentes) = 3/9$ (ou 1/3). Caso observe que a primeira prole teve dois descendentes, qual é a probabilidade de o besouro produzir 4 descendentes na próxima (totalizando 6)? Intuitivamente, parece que essa probabilidade também deveria ser de 1/3, pois há 3 resultados em F e somente um deles (2,4) resulta em 6 descendentes ao total. Esta é a resposta correta. Mas por que a probabilidade não é igual a 1/9, que é a de obter um (2,4) dentro do conjunto *Aptitão*? A resposta rápida é que a informação adicional de saber o tamanho da primeira prole influencia a probabilidade do resultado reprodutivo total. Probabilidades condicionais resolvem o enigma.

Probabilidades condicionais

Se estivermos calculando a probabilidade de um evento complexo e tivermos informação sobre um resultado daquele evento, por consequência devemos modificar as estimativas das probabilidades dos outros resultados. As estimativas atualizadas são chamadas **probabilidades condicionais** e as escrevemos como:

$$P(A \mid B)$$

ou a probabilidade do evento ou resultado A *dado* o evento ou resultado B. A barra vertical (|) indica que a probabilidade de A é calculada assumindo que o evento ou resultado B já ocorreu. A probabilidade condicional é definida como:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1.6}$$

Tal definição deveria fazer sentido intuitivo. Se o resultado B já ocorreu, então qualquer resultado no espaço amostral original não contido em B (i.e., B^c) não pode ocorrer, motivo pelo qual restringimos os resultados de A que poderíamos observar àqueles que também ocorrem em B. Assim, $P(A \mid B)$ deve, de alguma forma, estar relacionado a $P(A \cap B)$ e a definimos como sendo proporcional àquela interseção. O denominador $P(B)$ é o espaço amostral de eventos restringidos para os quais B já ocorreu. No exemplo do besouro girinídeo, $P(F \cap S) = 1/9$, e $P(F) = 1/3$, dividindo o primeiro pelo último obtivemos o valor para $P(F \mid S) = (1/9)/(1/3) = 1/3$, como o sugerido intuitivamente.

Rearranjando a fórmula da Equação 1.6 para calcular a probabilidade condicional, obtemos uma fórmula geral para calcular a probabilidade de uma interseção:

$$P(A \cap B) = P(A \mid B) \times P(B) = P(B \mid A) \times P(A) \qquad (1.7)$$

Você deveria relembrar que anteriormente, neste capítulo, definimos a probabilidade da interseção de dois eventos independentes como sendo igual a $P(A) \times P(B)$. Este é um caso especial da fórmula para calcular a probabilidade de interseção usando a probabilidade $P(A \mid B)$. Simplesmente note que, se dois eventos, A e B, são independentes, então $P(A \mid B) = P(A)$, desta forma $P(A \mid B) \times P(B) = P(A) \times P(B)$. Pela mesma razão, $P(A \cap B) = P(B \mid A) \times P(A)$.

TEOREMA DE BAYES

Até agora discutimos probabilidades usando o que é conhecido como **paradigma frequentista**, no qual as probabilidades são estimadas como as frequências relativas dos resultados, baseadas em um conjunto de observações infinitamente grande. Cada vez que cientistas desejam estimar a probabilidade de um fenômeno, começam assumindo a falta de conhecimento prévio sobre a probabilidade de um evento, e re-estimam a probabilidade com base em um grande número de observações. Em contraste, o **paradigma bayesiano**, fundamentado em uma fórmula para probabilidades condicionais desenvolvida por Thomas Bayes,[16] se estrutura na ideia de que os investigadores podem já possuir palpites das probabilidades de um evento, antes que as observações sejam conduzidas.

Thomas Bayes

[16] Thomas Bayes (1702-1761) foi um ministro não conformista da Capela Presbiteriana em Tunbridge Wells, sul de Londres, Inglaterra. Ele é mais conhecido por seu "*Essay towards solving a problem in the doctrine of chances*," publicado dois anos após sua morte no *Philosophical Transactions* 53:370-418 (1763). Bayes foi eleito membro da Sociedade Real de Londres (*Royal Society of London*) em 1742. Apesar de ser citado por suas contribuições à matemática, ele nunca publicou um artigo sobre a matéria em sua vida.

Tais **probabilidades *a priori*** podem ser baseadas em experiência prévia (que pode ser uma do tipo frequentista, estimada em estudos anteriores), intuição ou predições de modelos. Essas são então modificadas pelos dados das observações atuais para produzir **probabilidades *a posteriori*** (discutidas em mais detalhes no Capítulo 5). Contudo, mesmo no paradigma bayesiano, estimativas quantitativas de probabilidades *a priori* devem vir fundamentalmente de observações e experimentos.

A fórmula de Bayes para calcular probabilidades condicionais, agora conhecida como **Teorema de Bayes**, é:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} \qquad (1.8)$$

Tal fórmula é obtida por substituição simples. A partir da Equação 1.6, podemos escrever a probabilidade condicional $P(A \mid B)$ no lado direito da equação como:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

A partir da Equação 1.7, o numerador dessa segunda equação pode ser re-escrito como $P(B \mid A) \times P(A)$, produzindo o numerador do Teorema de Bayes. Pelo Axioma 1, o denominador, $P(B)$, pode ser re-escrito como $P(B \cap A) + P(B \cap A^c)$, pois esses dois termos se igualam a $P(B)$. Novamente, usando a fórmula para a probabilidade de uma interseção (Equação 1.7), esta soma pode ser re-escrita como $P(B \mid A) \times P(A) + P(B \mid A^c) \times P(A^c)$, que é o denominador do Teorema de Bayes.

Apesar do Teorema de Bayes ser simplesmente uma expansão da definição de probabilidade condicional, ele contém uma ideia muito poderosa: a de que a probabilidade de um evento ou resultado A, condicionado por outro evento B, pode ser determinada se você souber a probabilidade do evento B condicional ao evento A e o complemento de A, A^c (por isso o Teorema de Bayes é frequentemente chamado de teorema de probabilidade inversa). Iremos retornar a uma exploração detalhada do Teorema de Bayes e seu uso em inferência estatística no Capítulo 5.

Por enquanto, concluímos destacando a distinção muito importante entre $P(A \mid B)$ e $P(B \mid A)$. Apesar de essas duas probabilidades condicionais parecerem similares, elas medem coisas completamente diferentes. Como exemplo, relembre o caso das paineirinhas suscetíveis e resistentes e das lagartas.

Primeiro, considere a probabilidade condicional:

$$P(L \mid R)$$

Esta expressão indica a probabilidade de lagartas (L) serem encontradas *dada* a resistência de uma população de paineirinhas (R). Para estimar $P(L \mid R)$, precisa-

mos examinar uma amostra aleatória de populações de paineirinhas resistentes e determinar a frequência com que estas populações possuem lagartas. Agora, considere a probabilidade condicional:

$$P(R \mid L)$$

Em contraste à expressão anterior, $P(R \mid L)$ é a probabilidade de que uma população de paineirinhas seja resistente (R), dado que ela esteja sendo consumida por lagartas (L). Para estimar $P(R \mid L)$, precisaríamos examinar uma amostra aleatória de lagartas e determinar a frequência com a qual elas estavam de fato se alimentando em populações de paineirinhas resistentes.

Essas são duas quantidades bastante diferentes e suas probabilidades são determinadas por diferentes fatores. A primeira probabilidade condicional $P(L \mid R)$ (a probabilidade da ocorrência de lagartas em razão da resistência das paineirinhas) irá depender, entre outras coisas, da extensão a que resistência a herbívora é direta ou indiretamente responsável pela ocorrência de lagartas. Em contraste, a segunda probabilidade condicional $P(R \mid L)$ (a probabilidade de resistência de populações de paineirinhas em razão da presença de lagartas) irá depender, em parte, da incidência de lagartas em plantas resistentes *versus* outras situações que poderiam resultar na presença de lagartas (p. ex., lagartas se alimentando em outras espécies de plantas ou plantas suscetíveis que ainda não morreram).

Por fim, perceba que probabilidades condicionais se distinguem das simples: $P(L)$ é somente a probabilidade de que um indivíduo escolhido de maneira aleatória seja uma lagarta, e $P(R)$ é a probabilidade de que uma paineirinha escolhida ao acaso seja resistente a lagartas. No Capítulo 5, veremos como usar o Teorema de Bayes para calcular probabilidades condicionais quando não podemos medir diretamente $P(A \mid B)$.

RESUMO

A probabilidade de um resultado é apenas o número de vezes que um resultado ocorre dividido pelo total de observações. Se probabilidades simples são conhecidas ou estimadas, as de eventos complexos (Evento *A ou* Evento *B*) podem ser determinadas por meio da soma; probabilidades de eventos compartilhados (Evento *A e* Evento *B*) podem ser determinadas por meio da multiplicação. A definição de probabilidade, juntamente aos axiomas de aditividade de probabilidades e às três operações com conjuntos (união, interseção e complemento), forma os fundamentos do **cálculo de probabilidades**.

2

Variáveis Aleatórias e Distribuições de Probabilidades

No Capítulo 1 exploramos a noção de probabilidade e a ideia de que o resultado de uma única observação é um evento incerto. Contudo, quando acumulamos dados de muitas observações, começamos a ver padrões regulares na frequência de distribuição dos eventos (p.ex., Figura 1.1). Neste capítulo exploraremos algumas funções matemáticas úteis que podem gerar tais distribuições de frequência.

Certas distribuições de probabilidade são assumidas por muitos métodos estatísticos. Por exemplo, a análise de variância paramétrica (ANOVA, *ver* Capítulo 10) assume que os valores medidos em amostras aleatórias se ajustam à distribuição normal, ou distribuição com "forma de sino", e que a variância da distribuição seja similar entre os diferentes grupos. Se os dados se ajustam a estas premissas, a ANOVA pode ser utilizada para testar diferenças entre as médias de grupos. Distribuições de probabilidade também podem ser utilizadas para construir modelos e fazer predições. Por fim, distribuições de probabilidade podem se ajustar a conjuntos de dados reais sem que se especifique um modelo mecanístico em particular. Usamos distribuições de probabilidades porque elas funcionam – ajustam-se a muitos dados no mundo real.

Nossa primeira incursão na teoria de probabilidades envolveu a estimativa da probabilidade de resultados individuais, como a captura de presas por plantas jarro ou a reprodução de besouros carapeta. Encontramos valores numéricos para cada uma dessas probabilidades usando regras simples, ou **funções**. Mais formalmente, a regra matemática (ou função) que atribui um dado valor numérico a cada resultado possível de um experimento no espaço amostral de interesse é chamada de **variável aleatória**. Usamos o termo "variável aleatória" no sentido matemático, não no sentido coloquial de um evento aleatório.

Surgem em duas formas: **variáveis aleatórias discretas** e **variáveis aleatórias contínuas**. A primeira tem valores finitos ou contáveis (como os números

inteiros; ver a nota de rodapé 4, Capítulo 1). Exemplos comuns incluem presença ou ausência de uma dada espécie (que recebe os valores de 0 ou 1), ou o número de descendentes, de folhas ou de pernas (valores inteiros). O segundo tipo, por outro lado, pode assumir qualquer valor dentro de um intervalo regular. Exemplos incluem a biomassa de um estorninho, a área da folha consumida por um herbívoro ou o conteúdo de oxigênio dissolvido em uma amostra de água. Por um lado, todos os valores de uma variável contínua que medimos no mundo real são discretos: nossa instrumentação apenas permite medirmos coisas a um nível finito de precisão.[1] Mas não há um limite teórico à precisão que podemos obter na medição de uma variável contínua. No Capítulo 7 retornaremos à distinção entre as variáveis discreta e contínua como uma consideração-chave para os delineamentos de experimentos de campo e amostral.

VARIÁVEIS ALEATÓRIAS DISCRETAS

Variáveis aleatórias de Bernoulli

O experimento mais simples possui apenas dois resultados, como a presença ou ausência de organismos, moedas caírem em cara ou coroa, ou besouros se reproduzirem ou não. Uma variável aleatória que descreve o resultado de tal experimento é a **de Bernoulli**; um experimento com observações independentes no qual há apenas dois resultados possíveis para cada observação é uma **tentativa de Bernoulli.**[2] Usamos a notação:

[1] É importante entender a distinção entre **precisão** e **acurácia** em medidas. A acurácia se refere a quão próximo a medida está do seu valor real. Medidas acuradas são **destendenciadas**, indicando que elas não estão consistentemente nem abaixo nem acima do valor real. A precisão se refere à concordância entre uma série de medidas e ao grau com que essas medidas podem ser discriminadas. Por exemplo, uma medida pode ser precisa em 3 casas decimais, indicando que podemos discriminar as medidas após as 3 casas decimais. A acurácia é mais importante que a precisão. Seria muito melhor usar uma balança acurada que foi precisa em somente uma casa decimal que uma inacurada que foi precisa a cinco casas decimais. As casas decimais extras não lhe trazem mais próximo do valor real se seu instrumento é imperfeito ou tendencioso.

Jacob Bernoulli

[2] Jacob (vulgo Jaques ou James) Bernoulli (1654-1705) pertenceu a uma família famosa de físicos e matemáticos. Ele e seu irmão Johann são considerados atrás apenas de Newton em seu desenvolvimento do cálculo, mas arguiam constantemente acerca dos méritos relativos do trabalho um do outro. O maior trabalho matemático de Jacob, *Ars conjectandi* (publicado em 1713, 8 anos após sua morte), foi o primeiro grande texto sobre probabilidade. Ele incluiu a primeira exposição de muitos dos tópicos discutidos nos Capítulos 2 e 3, como teorias gerais de permutações e combinações, as primeiras provas do teorema binomial e a Lei dos Grandes Números. Jacob Bernoulli também fez contribuições substanciais à astronomia e à mecânica (física).

$$X \sim \text{Bernoulli}(p) \qquad (2.1)$$

para indicar que a variável X é uma de Bernoulli. O símbolo ~ é lido como "é distribuído como". Assim, X assume os valores do número de "sucessos" na observação (p. ex., presente, capturado, reproduziu), e p é a probabilidade de um resultado de sucesso. O exemplo mais comum de uma tentativa de Bernoulli é o arremesso de uma única moeda honesta (talvez não do Euro belga; *ver* a nota de rodapé 7, Capítulo 1, e o Capítulo 11), no qual a probabilidade de resultar cara = a probabilidade de dar coroa = 0,5. Contudo, mesmo uma variável com um grande número de resultados pode ser redefinida como uma tentativa de Bernoulli se for possível colapsar a amplitude de respostas a dois resultados. Por exemplo, um conjunto de 10 eventos reprodutivos do besouro carapeta pode ser analisado como uma única tentativa de Bernoulli, caso em que o sucesso é definido como *resultando de exatamente 6 descendentes* e o insucesso, como *resultando com mais ou menos de 6 descendentes.*

Exemplo de uma tentativa de Bernoulli

O censo de uma espécie de planta rara, *Rhexia mariana* (*meadow beauty*), em todos os municípios de Massachusetts fornece um exemplo de Tentativa de Bernoulli. A ocorrência de *Rhexia* é uma variável aleatória de Bernoulli X, que assume os valores de dois resultados: $X = 1$ (*Rhexia* presente) ou $X = 2$ (*Rhexia* ausente). Existem 349 municípios em Massachusetts, portanto, essa Tentativa de Bernoulli é a busca em todos esses municípios pela ocorrência de *Rhexia*. Devido ao fato de a planta ser rara, imagine que a probabilidade de que a *Rhexia* esteja presente (i. e., $X = 1$) é baixa: $P(X = 1) = p = 0{,}02$. Assim, se fizermos o censo em apenas um município, existem apenas 2% de chance ($p = 0{,}02$) de encontrar a *Rhexia*. Mas qual é a probabilidade esperada para que a *Rhexia* ocorra em 10 municípios, e não nos 339 municípios remanescentes?

Devido à probabilidade de a *Rhexia* estar presente em qualquer município ser $p = 0{,}02$, sabemos pelo Primeiro Axioma da Probabilidade (Capítulo 1) que a probabilidade da *Rhexia* estar ausente em qualquer município é $= (1 - p) = 0{,}98$. Por definição, cada evento (ocorrência em um município) nesta Tentativa de Bernoulli deve ser independente. Para tanto, assumimos que a presença ou ausência de *Rhexia* em um dado município independe em qualquer outro. Visto que a probabilidade de a *Rhexia* estar presente em um município é de 0,02 e a de que ocorra em outro é a mesma, a probabilidade que ela ocorra nesses dois municípios específicos é de $0{,}02 \times 0{,}02 = 0{,}0004$. Por extensão, a probabilidade de que a *Rhexia* ocorra em 10 municípios específicos é de $0{,}02^{10} = 1{,}024 \times 10^{-17}$ (ou 0,00000000000000001024, que é um número muito pequeno).

Contudo, este cálculo não nos dá a exata resposta que estamos buscando. Mais precisamente, queremos a probabilidade de que a *Rhexia* ocorra em exatamente

10 municípios *e* que *não* ocorra nos demais 339. Imagine que o primeiro município pesquisado não contenha *Rhexia*, o que deveria acontecer a uma probabilidade de $(1 - p) = 0{,}98$. O segundo município contém *Rhexia*, o que deveria acontecer a uma probabilidade de $p = 0{,}02$. Como a presença ou a ausência de *Rhexia* em cada município é independente, a probabilidade de ter um município com *Rhexia* e outro sem, após escolher dois municípios ao acaso, deve ser o produto de p e $(1 - p) = 0{,}02 \times 0{,}98 = 0{,}0196$ (*ver* Capítulo 1). Por extensão, consequentemente, a probabilidade de a *Rhexia* ocorrer em um dado conjunto de 10 municípios em Massachusetts deve ser $(0{,}02^{10})$ multiplicada pela de que a *Rhexia* não ocorra em nenhum dos outros 339 municípios $(0{,}98^{339}) = 1{,}11 \times 10^{-20}$. Mais uma vez, este é um número muito pequeno.

Porém, nosso cálculo, até agora, nos dá apenas a probabilidade de que a *Rhexia* ocorra em exatamente 10 municípios em particular (e em nenhum outro). Mas estamos de fato interessados na probabilidade de que a *Rhexia* ocorra em *quaisquer* 10 municípios de Massachusetts, não em uma lista específica de 10 municípios. A partir do Segundo Axioma da Probabilidade (Capítulo 1) aprendemos que as probabilidades de eventos complexos que ocorrem por diferentes vias podem ser calculadas somando-se as probabilidades de cada uma das vias. Desta forma, quantos conjuntos diferentes, com 10 municípios cada, são possíveis em uma lista de 349? Muitos! De fato, são possíveis $6{,}5 \times 10^{18}$ diferentes **combinações** de 10 com um conjunto de 349 (explicaremos este cálculo na próxima sessão). Portanto, a probabilidade de a *Rhexia* ocorrer em *quaisquer* 10 municípios é igual à de que ela ocorra em um conjunto de 10 municípios $(0{,}02^{10})$ multiplicada pela de que ela não ocorra nos 339 municípios remanescentes $(0{,}98^{339})$ multiplicada pelo número de combinações únicas que podem ser produzidas com um conjunto de 349 municípios $(6{,}5 \times 10^{18})$. Neste produto final = 0,07, há uma chance de 7% de encontrar exatamente 10 municípios com *Rhexia*. Este também é um número pequeno, mas não tão reduzido quanto os dois primeiros valores de probabilidade que havíamos calculado.

Muitas tentativas de Bernoulli = uma variável aleatória binomial

Como uma característica central da ciência é a replicação, raras vezes iremos conduzir apenas uma tentativa de Bernoulli. Pelo contrário, conduziríamos réplicas independentes de tentativas de Bernoulli em um único experimento. Definimos uma **variável aleatória binomial** X como sendo o número de resultados de sucesso em n tentativas de Bernoulli independentes. A notação para uma variável aleatória binomial é:

$$X \sim \text{Bin}(n,p) \qquad (2.2)$$

para indicar a probabilidade de obter X resultados de sucesso em n tentativas de Bernoulli, onde a probabilidade de um resultado de sucesso em qualquer evento é

p. Note que se $n = 1$, a variável aleatória binomial X é equivalente a uma variável aleatória de Bernoulli. Uma variável aleatória binomial é um dos tipos mais comuns de variáveis aleatórias encontradas em estudos ambientais e de ecologia. A probabilidade de obter X sucessos para uma variável aleatória binomial é:

$$P(X) = \frac{n!}{X!(n-X)!} p^X (1-p)^{n-X} \qquad (2.3)$$

onde n é o número de tentativas, X é o número de resultados de sucesso ($X \leq n$), e $n!$ significa n **fatorial**,[3] assim calculado:

$$n \times (n-1) \times (n-2) \times \ldots \times (3) \times (2) \times (1)$$

A Equação 2.3 tem três componentes, dos quais dois deles parecem familiares à nossa análise do problema que envolve a *Rhexia*. O componente p^X é a probabilidade de obter X sucessos independentes, cada um com probabilidade p. O componente $(1-p)^{(n-X)}$ é a probabilidade de obter $(n-X)$ insucessos, cada um a uma probabilidade de $(1-p)$. Note que a soma de sucessos (X) e o número de insucessos $(n-X)$ é n, o número total de tentativas de Bernoulli. Como vimos no exemplo da *Rhexia*, a probabilidade de obter X sucessos com probabilidade p e $(n-X)$ insucessos com probabilidade de $(1-p)$ é o produto desses dois eventos independentes $p^X(1-p)^{(n-X)}$.

Então por que precisamos do termo:

$$\frac{n!}{X!(n-X)!}$$

e de onde ele vem? A notação equivalente para esse termo é:

$$\binom{n}{X}$$

(lê-se "n sobre X"), e é conhecida como **coeficiente binomial**. O coeficiente binomial é necessário porque há mais de uma via para obter a maioria das combinações de sucessos e insucessos (as combinações de 10 municípios que descrevemos no exemplo anterior). Por exemplo, o resultado *um sucesso* em um conjunto de duas tentativas de Bernoulli pode, na verdade, ocorrer de duas formas: (1, 0) ou (0, 1). Dessa forma, a probabilidade de obter um sucesso em um conjunto de duas tentativas de Bernoulli é igual à probabilidade de obter um resultado com um sucesso $[= p(1-p)]$ vezes o número de resultados possíveis com um sucesso (= 2). Poderíamos escrever todos os resultados diferentes com X sucessos e contá-los,

[3] A operação de fatorial pode ser aplicada somente a números inteiros não negativos. Por definição, 0! = 1.

mas conforme o n fica grande, o X também fica; existem 2^n resultados possíveis para n tentativas.

Seria mais simples computar diretamente o número de resultados de X (sucessos), e isso é o que o coeficiente binomial faz. Retornando ao exemplo da *Rhexia*, temos 10 ocorrências dessa planta ($X = 10$) e 349 municípios ($n = 349$), porém não sabemos em quais municípios em particular ocorre. No começo de nossa busca, existiam inicialmente 349 municípios nos quais a *Rhexia* podia ocorrer. Uma vez que uma *Rhexia* seja encontrada em um dado município, em apenas outros 348 ela pode ocorrer. Assim, existem 349 combinações de municípios nas quais poderia haver apenas uma ocorrência e 348 combinações nas quais poderia haver uma segunda ocorrência. Por extensão, para X ocorrências, o número total de vias pelas quais as 10 ocorrências da *Rhexia* podem estar distribuídas nos 349 municípios deve ser

$$349 \times 348 \times 347 \times 346 \times 345 \times 344 \times 343 \times 342 \times 341 \times 340 = 2{,}35 \times 10^{25}$$

Em geral, o número total de formas para obter X sucessos em n tentativas é:

$$n \times (n-1) \times (n-2) \times \ldots \times (n-X+1)$$

o que parece muito com a fórmula do $n!$. Os termos na equação acima que faltam em $n!$ são todos os que vêm após $(n-X+1)$, ou

$$(n-X) \times (n-X-1) \times (n-X-2) \times \ldots \times 1$$

que é simplesmente igual a $(n-X)!$. Assim, se dividirmos $n!$ por $(n-X)!$, teremos o número total de formas para obter X sucessos em n tentativas. Mas isso não é exatamente o mesmo que o coeficiente binomial descrito acima, que além disso divide o resultado por $X!$:

$$\frac{n!}{(n-X)!\,X!}$$

A razão para esta divisão adicional é que não queremos contar duas vezes padrões idênticos de sucessos que simplesmente ocorreram em uma ordem diferente. Retornando à *Rhexia*, se primeiro encontrarmos populações no município de Barnstable e depois em Chatham, poderíamos não querer contar esta forma com um resultado diferente em que primeiro encontramos populações em Chatham e depois em Barnstable. Pela mesma razão acima, existem exatamente $X!$ formas para reordenar um dado resultado. Assim, precisamos "descontar" (dividir por) do resultado $X!$. A utilidade dessas diferentes **permutações** será mais aparente no Capítulo 5, quando discutirmos os Métodos de Monte Carlo ao testar hipóteses.[4]

[4] Mais um exemplo para persuadir o leitor de que o coeficiente binomial funciona. Considere o seguinte conjunto de 5 espécies de peixes marinhos:

{(bodião), (gobio), (Maria-da-toca), (enguia), (peixe-donzela)}

Quantos pares únicos de peixes podem ser formados? Se listarmos todos, encontramos 10:

(bodião), (Maria-da-toca)
(bodião), (gobio)
(bodião), (enguia)

(*Continua*)

A distribuição binomial

Agora que temos uma função simples para calcular a probabilidade de uma variável aleatória binomial, que faremos com ela? No Capítulo 1, apresentamos o histograma, um tipo de gráfico utilizado para apresentar de maneira concisa o número de observações que resultaram em algo particular. Similarmente, podemos fazer um gráfico do número de variáveis aleatórias binomiais de uma série de tentativas de Bernoulli que resultaram em cada um dentre as possíveis respostas. Tal histograma é chamado de **distribuição binomial** e é gerado a partir de uma variável aleatória binomial (Equação 2.3).

Por exemplo, considere a variável aleatória de Bernoulli para a qual a probabilidade de sucesso = probabilidade de insucesso = 0,5 (como jogar moedas honestas). Nosso experimento consistirá em jogar uma moeda honesta 25 vezes ($n = 25$) e nosso resultado possível é dado pelo espaço amostral do *número de caras* = {0, 1, ..., 24, 25}. Cada resultado X_i é uma variável aleatória binomial e a probabilidade de cada resultado é obtido pela fórmula binomial, a qual iremos nos referir como **função de distribuição de probabilidade**, porque ela é a função (ou regra) que fornece o valor numérico (a probabilidade) de cada resultado dentro do espaço amostral.

Usando nossa fórmula para uma variável aleatória binomial, podemos tabular as probabilidades de qualquer resultado possível em 25 buscas por *Rhexia* (Tabela 2.1). Podemos representar essa tabela em um histograma (Figura 2.1), que pode ser interpretado de duas formas. Primeiro, os valores do eixo-*y* podem ser vistos como a probabilidade de obter uma dada variável aleatória X (p. ex., o número de ocorrências de *Rhexia*) em 25 observações, em que a probabilidade de uma ocorrência é de 0,5. Nesta interpretação, nos referimos à Figura 2.1 como uma **distribuição de probabilidade**. Por consequência, se somarmos os valores da Tabela 2.1, o resultado é exatamente 1,0 (incluindo o erro de arredondamento), pois eles definem o espaço amostral inteiro (Primeiro Axioma da Probabilidade). Segundo, podemos interpretar os valores do eixo-*y* como a frequência relativa esperada para cada variável aleatória em um grande número de experimentos, cada um com 25 réplicas. Essa definição de frequência relativa se equipara à definição formal constante no Capítulo 1. Ela também é base do termo **frequentista**, que descreve estatísticas baseadas na frequência relativa de um evento com base em um número infinitamente grande de tentativas.

---

[4] (*Continuação*)
(bodião), (peixe-donzela)
(Maria-da-toca), (gobio)
(Maria-da-toca), (enguia)
(Maria-da-toca), (peixe-donzela)
(gobio), (enguia)
(gobio), (peixe-donzela)
(enguia), (peixe-donzela)

Usando o coeficiente binomial, poderíamos usar um $n = 5$ e $X = 2$ e chegar ao mesmo número:

$$\binom{5}{2} = \frac{5!}{3!2!} = \frac{120}{6 \times 2} = 10$$

TABELA 2.1 Probabilidades binomiais para p = 0,5 com 25 tentativas

Número de sucessos (X)	Probabilidade de (X) em 25 tentativas ($P(X)$)
0	0,00000003
1	0,00000075
2	0,00000894
3	0,00006855
4	0,00037700
5	0,00158340
6	0,00527799
7	0,01432598
8	0,03223345
9	0,06088540
10	0,09741664
11	0,13284087
12	0,15498102
13	0,15498102
14	0,13284087
15	0,09741664
16	0,06088540
17	0,03223345
18	0,01432598
19	0,00527799
20	0,00158340
21	0,00037700
22	0,00006855
23	0,00000894
24	0,00000075
25	0,00000003
	S = 1,00000000

A primeira coluna mostra a amplitude total do número de possíveis sucessos em 25 tentativas (i. e., de 0 a 25). A segunda dá a probabilidade de obter exatamente aquele número de sucessos, calculada a partir da Equação 2.3. Considerando os erros de arredondamento, tais probabilidades somam 1,0 à área total sob a curva de probabilidades. Note que um binômio com probabilidade $p = 0,5$ dá uma distribuição simétrica perfeita de sucessos, centralizada em 12,5, que é a expectativa da distribuição (Figura 2.1).

A Figura 2.1 é **simétrica** – o lado esquerdo e o lado direito da distribuição são imagens espelho uma da outra.[5] Contudo, aquele resultado é uma proprieda-

[5] Se perguntar à maioria das pessoas quais as chances de obter 12 ou 13 caras em 25 lançamentos de uma moeda honesta, elas irão dizer que é cerca de 50%. Contudo, a probabilidade real de obter 12 caras é somente de 0,155 (Tabela 2.1), cerca de 15,5%. Por que este número é tão pequeno? A resposta é que a equação binomial dá a **probabilidade exata**: o valor 0,155 é a probabilidade de

(*continua*)

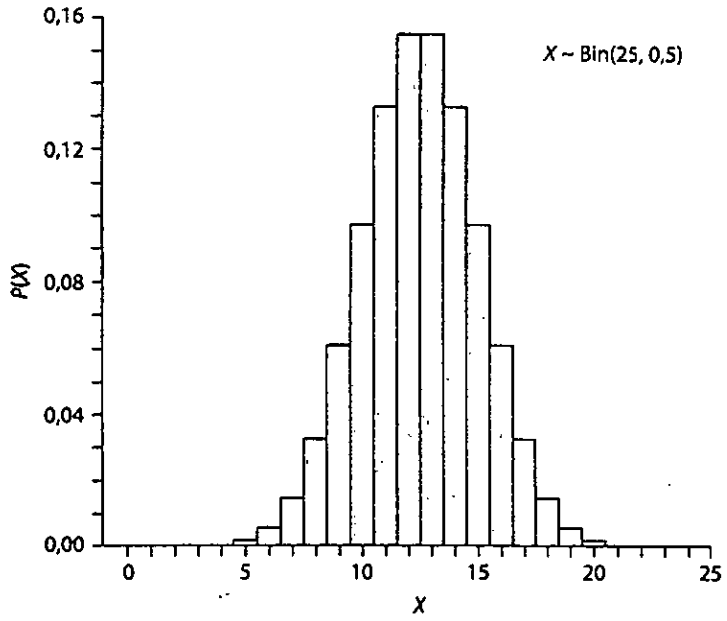


Figura 2.1 Distribuição de probabilidades para uma variável aleatória binomial. Há apenas dois resultados possíveis (p. ex., *Sim*, *Não*) em uma única tentativa. A variável *X* representa o número de resultados possíveis em 25 tentativas. A variável *P*(*X*) representa a probabilidade de obter determinado número de resultados positivos, calculada a partir da Equação 2.3. Como a probabilidade de um resultado de sucesso foi estipulada em $p = 0,5$, a distribuição binomial é simétrica, e o ponto no meio da distribuição, 12,5 (ou metade das 25 tentativas). Uma distribuição binomial é especificada por dois parâmetros: *n*, o número de tentativas; e *p*, a probabilidade de obter um resultado positivo.

de especial da distribuição binomial em que $p = (1 - p) = 0,50$. Alternativas são possíveis, como com moedas tendenciosas, que caem em cara com uma frequência superior a 50%. Por exemplo, com um $p = 0,80$ e 25 tentativas, uma forma diferente da distribuição binomial é obtida, como na Figura 2.2. Essa distribuição é assimétrica e deslocada para a direita; amostras com muitos sucessos são mais prováveis do que na Figura 2.1. Assim, a forma exata da distribuição binomial depende do número de tentativas (*n*), e da probabilidade de sucesso, (*p*). Você

[5] (*Continuação*)

obter *exatamente* 12 caras – nem mais nem menos. Contudo, cientistas estão frequentemente mais interessados em saber a probabilidade extrema ou **probabilidade de cauda**. Esta refere-se à chance de obter 12 ou menos caras em 25 tentativas e é calculada através da soma das probabilidades para cada resultado de 0 até 12 caras. Esta soma é de fato 0,50, e corresponde à área da metade esquerda da distribuição binomial da Figura 2.1.

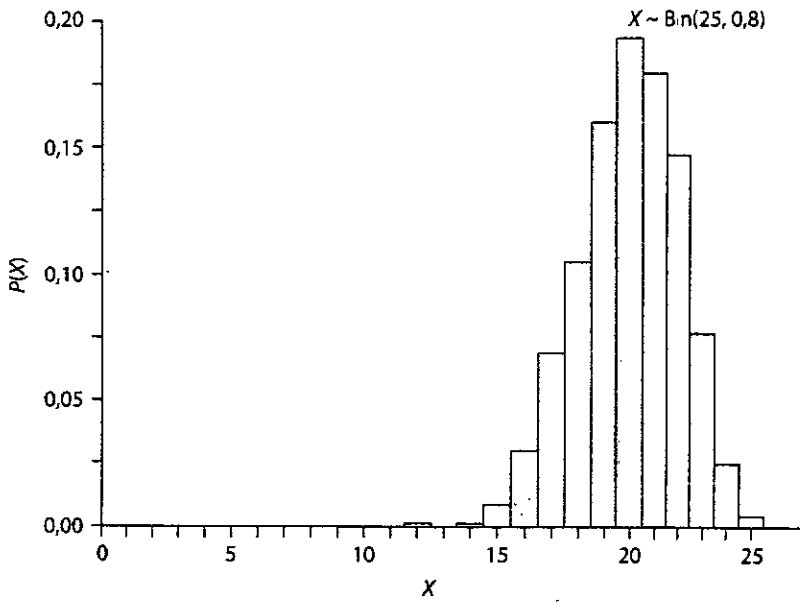


Figura 2.2 Distribuição de probabilidades para uma variável aleatória binomial $X \sim \text{Bin}(25, 0{,}8)$ – ou seja, em 25 tentativas e probabilidade de um resultado de sucesso para cada tentativa de $p = 0{,}8$. Esta figura segue o mesmo esboço e notação da Figura 2.1. Contudo, nesta, a probabilidade de um resultado positivo é $p = 0{,}8$ (em vez de $p = 0{,}5$), e a distribuição não é mais simétrica. Note que a expectativa para esta distribuição binomial é $25 \times 0{,}8 = 20$, que corresponde à moda (pico mais alto) do histograma.

pode tentar gerar suas próprias distribuições binomiais[6] usando valores diferentes para o n e para o p.

Variáveis aleatórias de Poisson

A distribuição binomial é apropriada para casos em que há um número fixo de tentativas (n) e a probabilidade de sucesso não é muito pequena. Contudo, a fórmula rapidamente se torna comprida e complicada quando o n se torna grande e o p pequeno, como a ocorrência de plantas ou animais raros. Não obstante,

[6] No Capítulo 1 (*ver* nota de rodapé 7), discutimos um experimento de girar-moedas com a nova moeda do Euro, que forneceu uma estimativa de $p = 0{,}56$ para $n = 250$ tentativas. Use planilhas ou programas estatísticos para gerar a distribuição binomial com esses parâmetros e compare-a à de uma moeda honesta ($p = 0{,}50$, $n = 250$). Você verá que essas distribuições de probabilidades de fato diferem, e que a com $p = 0{,}56$ é deslocada um pouco para a direita. Agora, faça o mesmo exercício com $n = 25$, como na Figura 2.1. Com um tamanho amostral relativamente pequeno, será virtualmente impossível distinguir as duas distribuições. Em geral, quanto mais próximas as expectativas de duas distribuições (*ver* Capítulo 3), maior será a amostra necessária para distingui-las.

a binomial somente é útil quando podemos contar diretamente as tentativas. Porém, em muitos casos, não contamos tentativas individuais, mas sim eventos que ocorrem em uma amostra. Por exemplo, suponha que as sementes de uma orquídea (uma planta de floração) sejam dispersas aleatoriamente em uma região, e que contemos o número de sementes dentro de quadrados amostrais de tamanho fixo. Nesse caso, a ocorrência de cada semente representa um resultado de sucesso de uma tentativa de dispersão que não foi observada. No entanto, realmente não podemos dizer quantas tentativas foram necessárias para gerar essa distribuição. Amostras no tempo também podem ser tratadas assim, como a contagem do número de pássaros que visitam um alimentador em um período de 30 minutos. Para tais distribuições, usamos a **distribuição de Poisson**, ao invés da distribuição binomial.

Uma **variável aleatória de Poisson**[7] X é o número de ocorrências de um evento registrado em uma amostra de área fixa ou durante um intervalo fixo de tempo. Variáveis aleatórias de Poisson são utilizadas quando tais ocorrências são raras – ou seja, quando o número mais comum de contagens em uma amostra é 0. A variável aleatória de Poisson, por si só, é o número de eventos em cada amostra. Como sempre, assumimos que as ocorrências de cada evento são independentes umas das outras.

Variáveis aleatórias de Poisson são descritas por um único parâmetro λ, algumas vezes referido como **razão de Poisson**, pois variáveis aleatórias de Poisson podem descrever a frequência de eventos raros no tempo. O parâmetro λ é o valor médio do número de ocorrências do evento em cada amostra (ou sobre cada intervalo de tempo). Estimativas de λ podem ser obtidas pela coleta de dados ou do conhecimento prévio. A notação para uma variável aleatória de Poisson é:

Siméon-Denis Poisson

[7] Variáveis aleatórias de Poisson são nomeadas para Siméon-Denis Poisson (1781-1840), que afirmou que "a vida foi boa apenas para duas coisas: para fazer matemática e para ensiná-la" (*fide* Boyer, 1968). Ele aplicou de matemática à física (em seu trabalho de dois volumes *Traité de mécanique*, publicado em 1811 e 1833, e realizou uma primeira derivação da Lei dos Grandes Números (*ver* Capítulo 3). Seu tratado em probabilidade, *Recherches sur la probabilité dés jugements*, foi publicado em 1837.

A mais famosa referência literária à distribuição de Poisson ocorre no romance de Thomas Pynchon *O arco-íris da gravidade* (*Gravity's rainbow*, 1972*). Um dos personagens principais do romance, Roger Mexico, trabalhava para a *The White Visitation* em *PISCES* (*Psychological Intelligence Scheme for Expediting Surrender*) durante a Segunda Guerra Mundial. Ele aponta, em um mapa de Londres, os pontos de ocorrência atingidos por bombas nazistas e ajusta os dados a uma distribuição de Poisson. *O arco-íris da gravidade*, que ganhou o prêmio National Book Award em 1974, contém diversas outras referências à estatística, matemática e ciência. *Ver* Simberloff (1978) para uma discussão da entropia e biofísica do romance de Pynchon.

* N. de T. Data da publicação do título em inglês. Em português, o livro foi publicado em 1998, pela Editora Companhia das Letras.

$$X \sim \text{Poisson}(\lambda) \tag{2.4}$$

e calculamos a probabilidade de qualquer observação x como:

$$P(x) = \frac{\lambda^x}{x!} e^{-\lambda} \tag{2.5}$$

onde e é uma constante, a base do logaritmo natural ($e \sim 2{,}71828$). Por exemplo, suponha que o número médio de plântulas de orquídeas encontrado em um quadrado de 1 m^2 é 0,75. Qual é a chance de um único quadrado conter 4 plântulas? Neste caso, $\lambda = 0{,}75$ e $x = 4$. Usando a Equação 2.5,

$$P\,(4\ plântulas) = \frac{0{,}75^4}{4!} e^{-0{,}75} = 0{,}0062$$

Um evento muito mais provável é um quadrado não conter nenhuma plântula:

$$P\,(0\ plântulas) = \frac{0{,}75^0}{0!} e^{-0{,}75} = 0{,}4724$$

Existe uma relação íntima entre as distribuições de Poisson e Binomial. Para uma variável aleatória binomial $X \sim \text{Bin}(n, p)$, na qual o número de sucessos k é muito pequeno em relação ao tamanho amostral (ou número de tentativas) n, podemos ter uma aproximativa de X usando a distribuição de Poisson, e estimando $P(X)$ como Poisson(λ). Contudo, a distribuição binomial depende da probabilidade de sucesso p e do número de tentativas n, enquanto a distribuição de Poisson depende apenas do número médio de eventos por amostra, λ.

Uma família inteira de distribuições de Poisson é possível, dependendo do valor de λ (Figura 2.3). Quando λ é pequeno, a distribuição forma um forte "J invertido" com os eventos mais prováveis sendo de 0 ou 1 ocorrência por amostra, mas com uma longa cauda de probabilidades que se estende à direita e as quais correspondem a amostras muito raras com muitos eventos. Conforme λ aumenta, o centro da distribuição se desloca para a direita, que se torna mais simétrica, assemelhando-se a uma distribuição normal ou binomial. As distribuições binomial e a de Poisson são discretas e possuem apenas números inteiros, com um valor mínimo de 0. Contudo, a distribuição binomial sempre é limitada por valores entre 0 e n (número de tentativas). Em contraste, a cauda do lado direito da distribuição de Poisson é ilimitada e se estende ao infinito, apesar de as probabilidades rapidamente se tornarem muito pequenas para números grandes de eventos em uma única amostra.

Um exemplo de variável aleatória de Poisson: distribuição de uma planta rara

Para fechar esta discussão, descrevemos a aplicação da distribuição de Poisson à da planta rara *Rhexia mariana* em Massachusetts. Os dados consistem do nú-

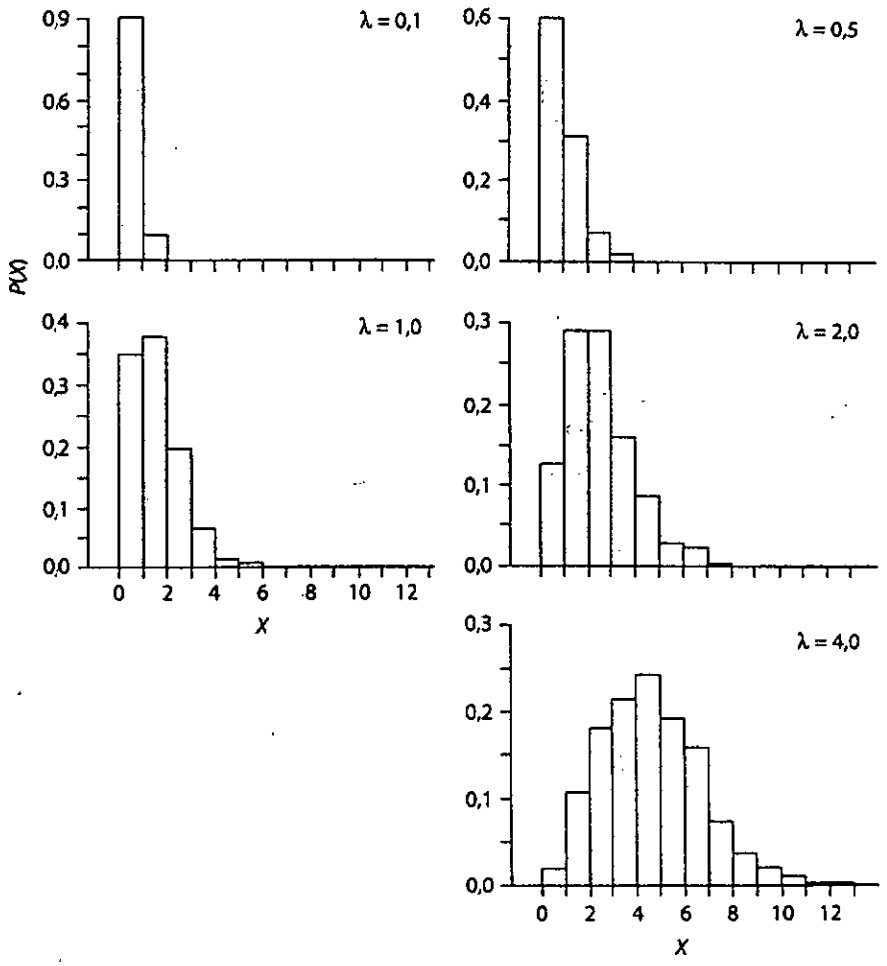


Figura 2.3 Famílias de curvas de Poisson. Cada histograma representa uma distribuição de probabilidades com diferente razão de Poisson λ. O valor de *X* indica o número de resultados positivos, e $p(X)$ é a probabilidade daquele resultado calculada através da Equação 2.5. Em contraste às distribuições binomiais mostradas nas Figuras 2.1 e 2.2, a de Poisson é especificada por um único parâmetro, λ, que é a taxa de ocorrência de eventos independentes (ou o número durante um dado intervalo de tempo). Conforme aumenta o valor do λ, a distribuição de Poisson se torna mais simétrica. Quando λ é muito grande, a distribuição de Poisson é praticamente indistinguível de uma normal.

mero de populações de *Rhexia* registrados em cada um dos 349 municípios de Massachusetts entre 1913 e 2001 (Craine, 2002). Apesar de cada população poder ser considerada uma tentativa de Bernoulli independente, nós sabemos apenas o número de populações, e não o de plantas individuais; os dados em cada município é nossa unidade amostral. Nessas unidades amostrais (municípios), a maioria

(344) não continha nenhuma população em nenhum dos anos, mas um dos municípios continha até 5 populações. O número médio de populações é 12/349, ou 0,03438 populações/município, que iremos usar como o parâmetro, λ, da razão da distribuição de Poisson. Posteriormente, usamos a Equação 2.5 para calcular a probabilidade de obter um dado número de populações em cada município. Multiplicando essas probabilidades pelo número do tamanho amostral (n = 349), obtemos a frequência esperada em cada classe de populações (de 0 a 5), que pode ser comparada ao número observado (Tabela 2.2).

As frequências observadas correspondem bem às predições do modelo de Poisson; elas não são significativamente diferentes uma da outra (usamos um teste de chi-quadrado para compará-las; *ver* Capítulo 11). Podemos imaginar que o município que tem cinco populações de uma planta rara seja um alvo-chave para as atividades de manejo e conservação – o solo é especialmente bom ou há hábitats únicos naquele município? Independente da causa final da ocorrência e persistência da *Rhexia* nesses municípios, o bom ajuste desses dados à distribuição de Poisson sugere que as populações são aleatórias e independentes; cinco populações em um único município podem apenas representar sorte. Portanto, estabelecer estratégias de manejo buscando por municípios com o mesmo conjunto de características físicas pode não resultar na conservação bem-sucedida da *Rhexia* naqueles municípios.

TABELA 2.2 Frequências esperadas e observadas do número de populações de uma planta rara (*Rhexia mariana*) em municípios de Massachusetts

Número de populações de *Rhexia*	Probabilidade de Poisson	Frequência de Poisson esperada	Frequência observada
0	0,96	337,2	344
1	0,033	11,6	3
2	$5{,}7 \times 10^{-4}$	0,19	0
3	$6{,}5 \times 10^{-6}$	$2{,}3 \times 10^{-3}$	0
4	$5{,}6 \times 10^{-8}$	$1{,}9 \times 10^{-5}$	1
5	$3{,}8 \times 10^{-10}$	$1{,}4 \times 10^{-7}$	1
Total	1,0000	349	349

A primeira coluna mostra quantas populações foram encontradas em um determinado número de municípios. Apesar de não haver nenhum máximo teórico para esse número, o observado nunca excedeu 5. Portanto, apenas os valores de 0 a 5 são mostrados. A segunda coluna apresenta a probabilidade de Poisson em obter o número observado, assumindo uma distribuição com razão de Poisson média de $\lambda = 0{,}03438$ populações/cidade. Tal valor corresponde à frequência observada nos dados (12 populações observadas divididas por 349 municípios investigados = 0,03438). A terceira coluna mostra a frequência de Poisson esperada, que simplesmente é a probabilidade de Poisson multiplicada pelo tamanho amostral de 349. Na última coluna, temos a frequência observada, ou seja, o número de municípios que contiveram certo número de populações de *Rhexia*. Note a proximidade entre a frequência observada e a frequência de Poisson esperada, sugerindo que a ocorrência de populações de *Rhexia* em um município é um evento aleatório independente. (Dados retirados de Craine, 2002.)

O valor esperado de uma variável aleatória discreta

Variáveis aleatórias assumem muitos valores diferentes, mas a distribuição inteira pode ser resumida determinando seu valor típico ou média aritmética. Você está familiarizado com a média aritmética como sendo uma medida da tendência central de um conjunto de números, que também será abordada no Capítulo 3, além de outras estatísticas descritivas úteis que podem ser calculadas a partir de dados. Contudo, quando estamos tratando com distribuições de probabilidade, a simples média aritmética é enganosa. Por exemplo, considere uma variável aleatória binomial X que pode assumir os valores 0 e 50 com probabilidades de 0,1 e 0,9, respectivamente. A média aritmética desses dois valores = (0 + 50)/2 = 25, mas o valor mais provável dessa variável aleatória é 50, que ocorre em 90% das vezes. Para obter o valor médio provável, calculamos a média aritmética dos valores, ponderada por sua probabilidade. Assim, os valores com probabilidades maiores contribuem mais que os com probabilidades menores.

Formalmente, considere uma variável aleatória discreta X, que pode assumir os valores $a_1, a_2, \ldots, a_n$, com probabilidades $p_1, p_2, \ldots, p_n$, respectivamente. Definimos o **valor esperado** de X, que escrevemos como $E(X)$ (lê-se "a **expectativa** de X"), como sendo:

$$E(X) = \sum_{i=1}^{n} a_i p_i = a_1 p_1 + a_2 p_2 + \ldots + a_n p_n \qquad (2.6)$$

Três pontos em relação $E(X)$ são dignos de nota. Primeiro, a série:

$$\sum_{i=1}^{n} a_i p_i$$

pode ter um número finito ou infinito de termos, dependendo da amplitude de valores que X pode assumir. Segundo, diferente de uma média aritmética, a soma não é dividida pelo número de termos. Como probabilidades são pesos relativos (elas são redimensionadas na escala entre 0 e 1), a divisão já foi realizada de maneira implícita. Por fim, exceto quando todos p_i são iguais, essa soma não é o mesmo que a média aritmética (*ver* Capítulo 3 para uma discussão adicional). Para as três variáveis aleatórias apresentadas – Bernoulli, binomial e Poisson – os valores esperados são p, np e λ, respectivamente.

A variância de uma variável aleatória discreta

A expectativa de uma variável aleatória descreve a média aritmética, ou tendência central dos valores. Contudo, este número (como todas as médias aritméticas) não dá nenhuma ideia sobre a dispersão, ou *variação*, entre os valores. E, como em uma família com número médio de 2,2 filhos, a expectativa não necessariamente dará uma representação acurada sobre um dado individual. De fato, para algumas

distribuições discretas, como a binomial ou a de Poisson, nenhuma das variáveis aleatórias poderá se igualar à expectativa.

Por exemplo, uma variável aleatória binomial que pode assumir os valores de -10 e +10, cada valor com probabilidade de 0,5, tem a mesma expectativa que uma variável aleatória binomial que pode assumir os valores de -1.000 e +1.000, cada um com probabilidade de 0,5. Ambos têm $E(X) = 0$, mas em nenhuma distribuição um valor aleatório de 0 será gerado. Além disso, os valores observados da primeira distribuição são muito mais próximos ao esperado do que os da segunda. Apesar de a expectativa descrever com acurácia o ponto central de uma distribuição, precisamos quantificar de alguma forma a dispersão dos valores em torno do ponto central.

A **variância** de uma variável aleatória, que escrevemos como $\sigma^2(X)$, é uma medida de quanto os valores reais diferem do esperado. A variância de uma variável aleatória X é definida como:

$$\begin{aligned}\sigma^2(X) &= E[X - E(X)]^2 \\ &= \sum_{i=1}^{n} p_i \left(a_i - \sum_{i=1}^{n} a_i p_i \right)^2 \end{aligned} \qquad (2.7)$$

Na Equação 2.6, $E(X)$ é o valor esperado de X, os a_i são os diferentes valores possíveis da variável X, e cada um ocorre a uma probabilidade p_i.

Para calcular a $\sigma^2(X)$, primeiro calculamos a $E(X)$, subtraímos esse valor de X, e então elevamos a diferença ao quadrado. Esta é uma medida básica de quanto cada valor de X difere da expectativa $E(X)$.[8] Como existem diversos valores possíveis de X para variáveis aleatórias (p. ex., dois valores possíveis para uma variável de Bernoulli, n valores possíveis em tentativas de uma variável aleatória binomial, e infinitamente muitos valores para uma variável aleatória de Poisson), repetimos esse passo de subtrair e elevar ao quadrado para cada valor possível de X. Por fim, calculamos a expectativa desses desvios quadrados seguindo o mesmo procedimento da Equação 2.6: cada desvio quadrado é ponderado por seu valor de probabilidade de ocorrência (p_i) e posteriormente são somados.

Assim, no exemplo acima, se Y é uma variável aleatória binomial que pode assumir os valores -10 e +10, cada um com $P(Y)$ 0,5, então $\sigma^2(Y) = 0{,}5(-10 - 0)^2 + 0{,}5(10 - 0)^2 = 100$. Similarmente, se Z é uma variável aleatória binomial que assume os valores −1.000 e +1.000, então $\sigma^2(Z) = 100.000$. Tal resultado apoia a

[8] Você pode estar se perguntando por que elevar o desvio ao quadrado $E(X)$ uma vez que ele é calculado. Se simplesmente somar os desvios sem elevar ao quadrado descobrirá que a soma é sempre 0, pois $E(X)$ é o ponto central de todos os valores de X. Estamos interessados na magnitude da diferença, independente do seu sinal, de forma que poderíamos usar os valores do desvio absoluto $|X - E(X)|$. Contudo, a álgebra de valores absolutos não é tão simples quanto elevar ao quadrado, que também tem melhores propriedades matemáticas. Essa soma de quadrados, $[E(X - E(X))]^2$, forma a base da análise de variância (*ver* nota de rodapé 1 e o Capítulo 10).

intuição de que Z tem uma "dispersão" maior que Y. Por fim, note que se todos os valores da variável aleatória são idênticos [i.e., $X = E(X)$, então $\sigma^2(X) = 0$, e não há variação nos dados.

Para as três variáveis discretas apresentadas – Bernoulli, binomial e Poisson – as variâncias são $p(1-p)$, $np(1-p)$ e λ, respectivamente (Tabela 2.3). A expectativa e a variância são os dois descritores mais importantes de uma variável aleatória ou de uma distribuição de probabilidade. O Capítulo 3 discute outras medidas descritivas de variáveis aleatórias. Agora, vamos voltar nossa atenção para as variáveis contínuas.

VARIÁVEIS ALEATÓRIAS CONTÍNUAS

Muitas variáveis ecológicas e ambientais não podem ser representadas como variáveis contínuas. Por exemplo, o comprimento das asas de aves ou a concentração de pesticidas em tecidos de peixes podem assumir qualquer valor (delimitados por um **intervalo** com limites superior e inferior apropriados) e a **precisão** com que o valor é medido é limitada apenas pela instrumentação disponível. Quando trabalhamos com variáveis aleatórias discretas, somos capazes de definir o espaço amostral total como sendo um conjunto de resultados discretos possível. Contudo, quando com variáveis contínuas, não podemos identificar todos os eventos ou resultados possíveis, pois existem milhares deles (frequentemente são muitos e incontáveis; *ver* nota de rodapé 4, Capítulo 1). Similarmente, como as observações podem assumir qualquer valor dentro de um intervalo

TABELA 2.3 Três distribuições estatísticas discretas

Distribuição	Valor da Probabilidade	$E(X)$	$\sigma^2(X)$	Comentários	Exemplo ecológico
Bernoulli	$P(X) = p$	p	$p(1-p)$	Use para resultados dicotômicos.	Se reproduzir ou não, esta é a questão.
Binomial	$P(X) = \binom{n}{X} p^X (1-p)^{n-X}$	np	$np(1-p)$	Use para o número de sucessos em n tentativas independentes.	Presença ou ausência de espécies.
Poisson	$P(x) = \frac{\lambda^x}{x!} e^{-\lambda}$	λ	λ	Use para eventos independentes raros, onde λ é a taxa na qual os eventos ocorrem no espaço ou no tempo.	Distribuição de espécies raras em uma paisagem.

As equações na coluna "valor da probabilidade" determinam a chance de se obter um valor particular X em cada distribuição. A expectativa $E(X)$ da distribuição de valores é estimada pela média ou média aritmética da amostra. A variância $\sigma^2(X)$ é uma medida de dispersão ou o desvio das observações do $E(X)$. Essas distribuições são para variáveis discretas, as quais são medidas com números inteiros ou contagens.

definido, é difícil definir a probabilidade de obter um valor em particular. Ilustramos esses assuntos conforme descrevemos nosso primeiro tipo de variável aleatória contínua, a do tipo uniforme.

Variáveis aleatórias uniformes

Ambos os problemas mencionados acima – definir o espaço amostral apropriado e obter a probabilidade para qualquer dado valor – são solucionáveis. Primeiramente, reconhecemos que o espaço amostral não é mais discreto, mas contínuo. Em um espaço amostral contínuo, não consideramos mais resultados discretos (como $X = 2$), mas mantemos o foco em eventos que ocorrem dentro de um determinado subintervalo (como $1,5 < X < 2,5$). A probabilidade de que um evento ocorra dentro de um subintervalo pode ser tratada como um evento, e as nossas regras de probabilidade continuam valendo. Começaremos com um exemplo teórico, o **intervalo unitário fechado**, que contém todos os números entre 0 e 1, incluindo-os, o qual escrevemos como [0, 1]. Neste intervalo unitário fechado, suponha que a probabilidade de um evento X ocorrer entre 0 e $1/4 = p_1$, e a probabilidade desse mesmo evento ocorrer entre 1/2 e $3/4 = p_2$. Pela regra em que probabilidade da união de dois eventos independentes é igual à soma de suas probabilidades (*ver* Capítulo 1), a chance de que X ocorra em qualquer um desses dois intervalos (0 a 1/4 ou 1/2 a 3/4) $= p_1 + p_2$.

A segunda regra importante é que todas as probabilidades de um evento X em um espaço amostral contínuo deve somar 1 (é o Primeiro Axioma da Probabilidade). Suponha que, dentro de um intervalo unitário fechado, todos os resultados possíveis possuem a mesma probabilidade (imagine, por exemplo, um dado com um número infinito de lados). Apesar de não podermos definir com precisão a probabilidade de se obter o valor 0,1 em uma jogada desse dado, podemos dividir esse intervalo em 10 partes com **intervalo aberto** e *mesmo comprimento* {[0, 1/10], [1/10, 2/10], ..., [9/10, 1]} e calcular a probabilidade de qualquer jogada cair dentro de um desses subintervalos. Como você pode imaginar, a probabilidade de uma jogada cair dentro de qualquer um desses subintervalos poderia ser 0,1, e a soma de todas as probabilidades desse conjunto $= 10 \times 0,1 = 1,0$.

A Figura 2.4 ilustra este princípio. Neste exemplo, o intervalo varia de 0 a 10 (no eixo-x). Desenhe uma linha em L paralela ao eixo-x com um interceptor-y de 0,1. Para qualquer subintervalo U, a probabilidade de, em uma única jogada, o dado cair naquele intervalo pode ser obtida calculando a *área* do retângulo que delimita o subintervalo. Esse retângulo tem comprimento igual ao tamanho da altura do subintervalo = 0,1. Se dividirmos o intervalo em u subintervalos iguais, a soma de todas as áreas desses subintervalos será igual à da área de todo o intervalo: 10 (comprimento) $\times$ 0,1 (altura) = 1,0. Este gráfico também ilustra que a probabilidade de qualquer resultado a em particular

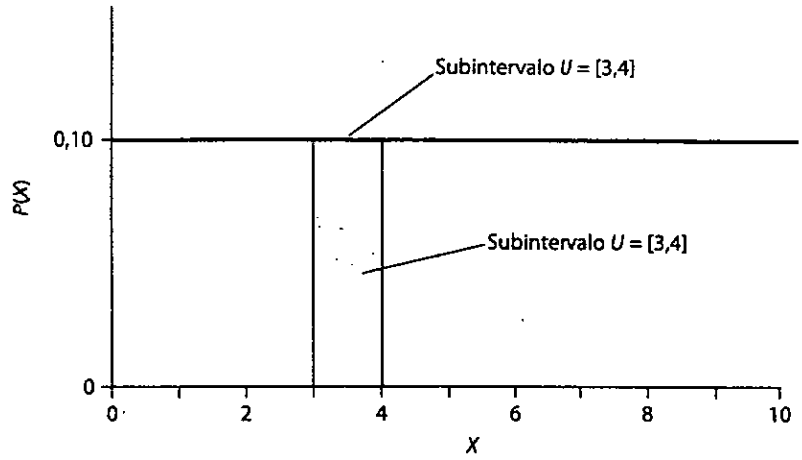


Figura 2.4 Distribuição uniforme com intervalo [0,10]. Em uma distribuição uniforme contínua, a probabilidade de um evento ocorrer em um subintervalo particular depende da área relativa do subintervalo; ela é a mesma independente de onde o subintervalo está inserido dentro dos limites da distribuição. Por exemplo, se a distribuição é delimitada por 0 e 10, a probabilidade de que um evento ocorra no subintervalo [3,4] é a área relativa delimitada por aquele subintervalo, que neste caso é 0,10. A probabilidade é a mesma para qualquer outro subintervalo com o mesmo tamanho, como [1,2] ou [4,5]. Se o subintervalo escolhido é maior, a probabilidade de um evento ocorrer naquele subintervalo será proporcionalmente maior. Por exemplo, a probabilidade de um evento ocorrer no subintervalo [3,5] é de 0,20 (desde que 2 dentre as 10 unidades do intervalo sejam transpassadas), e é de 0,6 para o subintervalo [2,8].

dentro de um espaço amostral contínuo é 0, pois o subintervalo que contém somente a é infinitamente pequeno, e um número infinitamente pequeno dividido por um número maior = 0.

Agora podemos definir uma **variável aleatória uniforme** X com respeito a qualquer intervalo I. A probabilidade de sua ocorrência em qualquer subintervalo U é igual ao produto $U \times I$. No exemplo ilustrado na Figura 2.4, definimos a seguinte função para descrever essa variável aleatória uniforme:

$$f(x) = \begin{cases} 1/10, & 0 \le x \le 10 \\ 0 \text{ caso contrário} \end{cases}$$

A função $f(x)$ é chamada de **função densidade de probabilidade (FDP)** para a distribuição uniforme. Em geral, a FDP de uma variável aleatória contínua é encontrada assinalando probabilidades com que variáveis aleatórias contínuas X ocorrem dentro de um intervalo I. A probabilidade de X ocorrer dentro de um intervalo I é igual à área da região delimitada por I no eixo-x e $f(x)$ no eixo-y. Pelas regras da probabilidade, a área total sob a curva descrita por FDP = 1.

Também podemos definir a **função de distribuição cumulativa** (FDC) de uma variável aleatória X com a função $F(y) = P(X < y)$. A relação entre FDP e FDC é a seguinte:

> Se X é uma variável aleatória com FDP $f(x)$, então FDC $F(y) = P(X < Y)$ é igual à área sob $f(x)$ no intervalo $x < y$.

A FDC representa a **probabilidade de cauda** – ou seja, a probabilidade de que uma variável X seja menor ou igual a algum valor y, $[P(X)<y]$ – e é o mesmo que o familiar valor de P que discutiremos no Capítulo 4. A Figura 2.5, abaixo nesta página, ilustra a FDP e a FDC para uma variável aleatória uniforme no intervalo unitário fechado.

O valor esperado de uma variável aleatória contínua

Apresentamos o valor esperado de uma variável aleatória no contexto de uma distribuição discreta. Para uma variável aleatória discreta,

$$E(X) = \sum_{i=1}^{n} a_i p_i$$

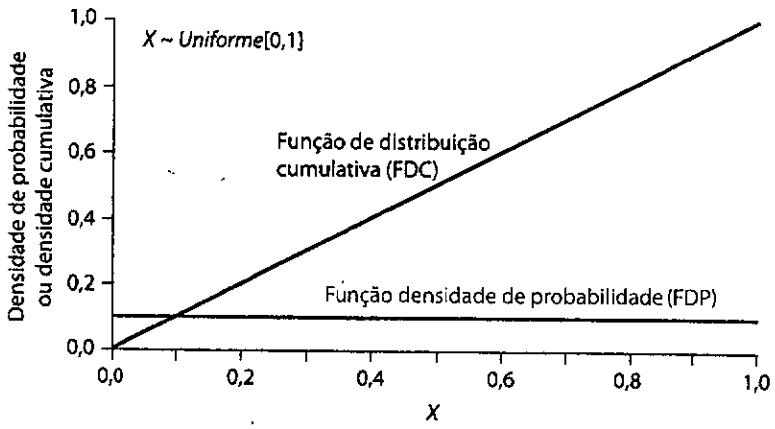


Figura 2.5 Função densidade de probabilidade e função de distribuição cumulativa para a distribuição uniforme medida sobre o intervalo unitário fechado [0,1]. A FDP mostra a probabilidade $P(X)$ para qualquer valor X. Nessa distribuição contínua, a probabilidade exata do valor X é tecnicamente 0, pois a área sob a curva é zero quando medida em um único ponto. Contudo, a área sob a curva de qualquer subintervalo mensurável é apenas a proporção da área total sob a curva, a qual por definição é 1,0. Na distribuição uniforme, a probabilidade de um evento em qualquer subintervalo é a mesma, indiferente de onde o intervalo está localizado. A FDC, para essa mesma distribuição, ilustra a área cumulativa sob a curva para o subintervalo que é delimitado inferiormente por 0 e superiormente por 1,0. Como esta é uma distribuição uniforme, tais probabilidades se acumulam de forma linear. Quando o fim do intervalo 1,0 é atingido, FDC = 1,0, pois a área inteira sob a curva foi incluída.

Este cálculo faz sentido, pois cada valor de a_i de X tem uma probabilidade associada p_i. Contudo, para distribuições contínuas, como a uniforme, a probabilidade de qualquer observação em particular é = 0, então usamos as probabilidades dos eventos ocorrerem dentro de subintervalos no espaço amostral. Usamos esta mesma abordagem para obter o valor esperado de uma variável aleatória contínua. Para encontrar o valor esperado de uma variável aleatória contínua usaremos subintervalos muito pequenos do espaço amostral, os quais denotaremos como Δx. Para uma função densidade de probabilidade $f(x)$, o produto de $f(x_i)$ e Δx resulta no valor da probabilidade de um evento ocorrer dentro do subintervalo Δx, escreve-se $P(X = x) = f(x_i)\Delta x$. Essa probabilidade é o mesmo que p_i no caso discreto. Como na Figura 2.4, o produto de $f(x_i)$ por Δx descreve a área de um retângulo muito estreito. No caso discreto, encontramos o valor esperado somando o produto de cada x_i pela sua probabilidade associada p_i. No contínuo, também encontramos o valor esperado de uma variável aleatória contínua somando os produtos de cada x_i e sua probabilidade $f(x_i)\Delta x$. Obviamente, o valor dessa soma dependerá do tamanho do subintervalo menor Δx. Contudo, se FDP $f(x)$ tem propriedades matemáticas "razoáveis" e se permitirmos que Δx fique cada vez menor, a soma

$$\sum_{i=1}^{n} x_i f(x_i)\Delta x$$

se aproxima de um valor único limitante. Esse valor limitante = $E(X)$ para uma variável aleatória contínua.[9] Para uma variável aleatória uniforme X, na qual $f(X)$ é definida sobre o intervalo $[a,b]$ e onde a < b,

$$E(X) = (b + a)/2$$

A variância de uma variável aleatória uniforme é:

$$\sigma^2(X) = \frac{(b-a)^2}{12}$$

Variável aleatória normal

Talvez a distribuição de probabilidades mais familiar seja a "curva de sino" – **distribuição de probabilidade normal** (ou **de gaussiana**). Essa distribuição forma as bases teóricas da regressão linear e da análise de variância (*ver* Capítulos 9 e 10) e se ajusta a muitos conjuntos de dados empíricos. Apresentaremos a distribuição

[9] Se você já estudou cálculo, irá reconhecer esta aproximação. Para uma variável aleatória contínua X, onde $f(X)$ é diferencial dentro do espaço amostral,

$$E(X) = \int xf(x)dx$$

A integral representa a soma do produto $x \times f(X)$, onde x se torna infinitamente pequeno em seu limite.

normal com um exemplo de aranhas da família Linyphiidae. Integrantes dos diferentes gêneros deste aracnídeo podem ser identificados pelo comprimento de seu espinho tibial. Um aracnólogo mediu 50 espinhos e obteve a distribuição ilustrada na Figura 2.6. O histograma dessas medidas ilustra vários traços característicos da distribuição normal.

Primeiro, note que a maioria das observações é agregada ao redor do centro, ou da média aritmética do comprimento do espinho tibial. Contudo, existem grandes caudas no histograma que se estendem para a direita e para a esquerda do centro. Em uma distribuição normal verdadeira (na qual medimos um número infinito de aranhas azaradas), estas caudas se estendem sem definição em ambas as direções, embora a densidade de probabilidade torna-se rápida e infimamente pequena conforme se afasta do centro. Em conjuntos de dados reais, as caudas não se estendem para sempre, pois temos uma quantidade limitada de dados e porque a maioria das variáveis medidas não pode assumir valores negativos. Por fim, note que a distribuição é aproximadamente simétrica: os lados direito e esquerdo do histograma são praticamente espelho um do outro.

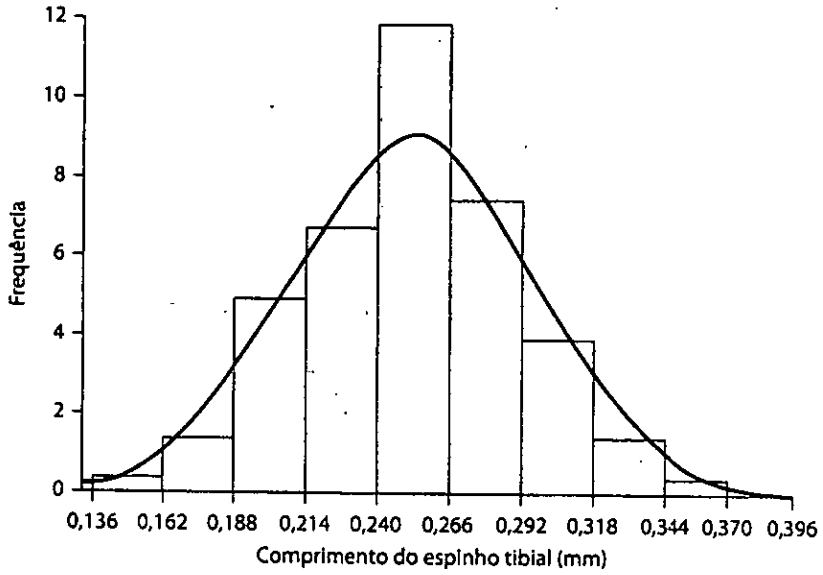


Figura 2.6 Distribuição normal de um conjunto de medidas morfológicas. Cada observação neste histograma representa uma de 50 medidas do comprimento do espinho tibial de aranhas (dados brutos na Tabela 3.1). As observações são agrupadas em "classes" que abrangem intervalos de 0,026-mm; o comprimento de cada barra é a frequência (o número de observações que caem naquela classe) Superposto ao histograma está a distribuição normal, com média de 0,253 e desvio-padrão de 0,0039. Apesar de o histograma não se acomodar perfeitamente à distribuição normal, o ajuste geral dos dados é muito bom: o histograma exibe um único pico central, uma distribuição aparentemente simétrica e um decréscimo estável na frequência das medidas muito grandes e muito pequenas nas caudas da distribuição.

Se considerarmos o comprimento tibial como uma variável aleatória X, podemos usar a FDP normal para aproximar essa distribuição. A distribuição normal é descrita por dois parâmetros, que chamaremos de μ e σ, portanto $f(X) = f(\mu,\sigma)$. A forma exata dessa função não é importante aqui,[10] mas tem as propriedades $E(X) = \mu, \sigma^2(X) = \sigma^2$, e a distribuição é simétrica em torno de μ. $E(X)$ é a expectativa e representa a tendência central dos dados, $\sigma^2(X)$ é a variância e representa a dispersão das observações em torno da expectativa. Uma variável X descrita por essa distribuição é chamada de **variável aleatória normal** ou de **variável aleatória gaussiana**,[11] e escreve-se:

[10] Se você almeja os detalhes, a FDP para a distribuição normal é:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

onde π é 3,14159..., e é a base do logaritmo natural (2,71828...), e μ e σ são **parâmetros** que definem a distribuição. Você pode ver a partir dessa fórmula que, quanto mais distante X está de μ, maior é o expoente negativo de e, e, portanto, menor é a probabilidade calculada de X. A FDC dessa distribuição,

$$F(X) = \int_{-\infty}^{X} f(x)dx$$

não possui uma solução analítica. A maioria dos textos estatísticos fornece tabelas para a FDC da distribuição normal. Ela também pode ser aproximada usando técnicas de integração numérica em programas de computador, como o Matlab ou o S-Plus.

Karl Friedrich Gauss

[11] A distribuição gaussiana é uma homenagem a Karl Friedrich Gauss (1777-1855), um dos mais importantes matemáticos na história. Um filho prodígio, ele tem a fama de ter corrigido um erro aritmético nos cálculos da folha de pagamentos de seu pai quando tinha apenas 3 anos. Gauss também provou o Teorema Fundamental da Álgebra (todo polinomial tem raiz com forma $a + bi$, onde $i = \sqrt{-1}$); o Teorema Fundamental da Aritmética (qualquer número natural pode ser representado com um produto de números primos único); formalizou a teoria dos números; e, em 1801, desenvolveu um método de ajustar uma linha a um conjunto de pontos usando mínimos quadrados (*ver* Capítulo 9). Infelizmente, Gauss não publicou seu método de mínimos quadrados, que geralmente é creditado a Legendre, que o publicou 10 anos depois. Gauss percebeu que a distribuição dos erros no ajuste das linhas por mínimos quadrados aproximava o que hoje chamamos de distribuição normal, que foi introduzida quase 100 anos antes pelo matemático americano Abraham de Moivre (1667-1754) em seu livro *The Doctrine of chances*. Como Gauss foi o mais famoso entre os dois (naquele tempo, os Estados Unidos não tinham destaque na matemática), a distribuição de probabilidade normal foi inicialmente tratada como distribuição gaussiana. Sozinho, De Moivre identificou a distribuição normal através de seus estudos da distribuição binomial descrita anteriormente neste capítulo. Contudo, o nome moderno "normal" não foi dado a essa distribuição até o fim do século dezenove (pelo matemático Poincaré), quando o estatístico Karl Pearson (1857-1936) redescobriu o trabalho de De Moivre e mostrou que sua descoberta acerca dessa distribuição precedeu Gauss.

$$X \sim N(\mu,\sigma) \tag{2.8}$$

Muitas distribuições normais diferentes podem ser criadas especificando-se valores diferentes de μ e σ. Contudo, estatísticos usam com frequência a **distribuição normal padrão**, em que $\mu = 0$ e $\sigma = 1$. A **variável aleatória normal padrão** associada é comumente tratada apenas como Z. $E(Z) = 0$, e $\sigma^2(Z) = 1$.

Propriedades úteis da distribuição normal

A distribuição normal possui três propriedades úteis. Primeiro, as distribuições normais podem ser adicionadas. Se você tem duas variáveis aleatórias normais independentes X e Y, a soma delas também é uma variável aleatória normal com $E(X + Y) = E(X) + E(Y)$ e $\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$.

Segundo, as distribuições normais podem ser facilmente transformadas com operações de **deslocamento** e **mudança de escala.** Considere duas variáveis aleatórias X e Y. Permita que $X \sim N(\mu,\sigma)$, e que $Y = aX + b$, onde a e b são constantes. Dizemos que a operação de multiplicar X pela constante a é uma operação de mudança de escala, porque uma unidade de X torna-se a unidades de Y; consequentemente, Y aumenta em função de a. Em contraste, a adição da constante b ao X é tratada como uma operação de deslocamento, pois simplesmente deslocamos a variável aleatória sobre b unidades ao longo do eixo-x através da adição de b a X. Operações de deslocamento e mudança de escala são ilustradas na Figura 2.7.

Se X é uma variável aleatória normal, então uma variável aleatória Y, criada por uma operação de mudança de escala, ou por uma operação de deslocamento, ou por ambas sobre a do tipo X, também é uma variável aleatória normal. Convenientemente, a expectativa e a variância da nova variável aleatória é uma função simples das constantes de deslocamento e escala. Para duas variáveis aleatórias $X \sim N(\mu,\sigma)$ e $Y = aX + b$, calculamos $E(Y) = a\mu + b$ e $\sigma^2(Y) = a^2\sigma^2$. Note que a expectativa da nova variável aleatória, $E(Y)$, é criada aplicando diretamente as operações de deslocamento e escala em $E(X)$. Contudo, a variância de $\sigma^2(Y)$ é alterada somente pela operação de escala. Intuitivamente, você pode observar que se todos os elementos de um conjunto de dados são deslocados pela adição de uma constante, a variância não deveria mudar, pois a dispersão relativa dos dados não foi afetada (*ver* Figura 2.7). Contudo, se aqueles dados forem multiplicados por uma constante a (mudança de escala), a variância será acrescida pela quantidade a^2; pois a distância relativa de cada elemento da expectativa agora é acrescida por um fator de a (Figura 2.7), esta quantidade é elevada ao quadrado no cálculo da variância.

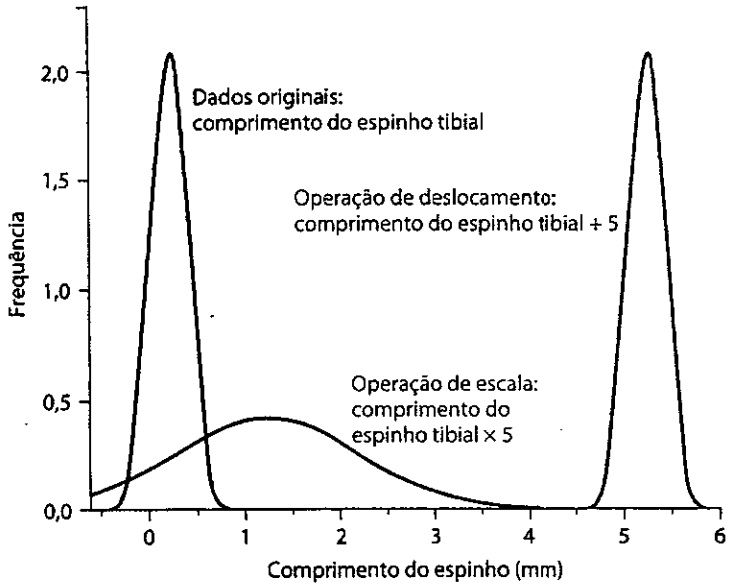


Figura 2.7 Operações de deslocamento e escala sobre uma distribuição normal. A distribuição normal possui duas propriedades algébricas convenientes. A primeira é uma operação de deslocamento: se a constante *b* é adicionada a um conjunto de medidas com média μ, a média da nova distribuição será deslocada para $\mu + b$, mas a variância não é afetada. A curva preta é o ajuste da distribuição normal a um conjunto de 200 medidas do comprimento do espinho tibial de aranhas (Figura 2.6). A curva cinza mostra a distribuição normal deslocada após o valor 5 ser acrescido a cada uma das observações originais. A média se deslocou 5 unidades para a direita, mas a variância não é alterada. Em uma operação de escala (curva verde), multiplicar cada observação por uma constante *a* causa um acréscimo na média por um fator de a^2. Esta curva é o ajuste da distribuição normal aos dados depois de terem sido multiplicados por 5. A média é deslocada para um valor 5 vezes maior que o original, e a variância aumenta por um fator de $5^2 = 25$.

Uma propriedade final (e conveniente) da distribuição normal é o caso especial de uma mudança de escala e operação de deslocamento em que $a = 1/\sigma$ e $b = -1(\mu/\sigma)$:

Para $X \sim N(\mu,\sigma)$, $Y = (1/\sigma)X - \mu/\sigma = (X - \mu)/\sigma$, obtemos

$$E(Y) = 0 \text{ e } \sigma^2(Y) = 1$$

que é uma variável aleatória normal padrão. Este é um resultado especialmente útil, pois significa que toda variável aleatória normal pode ser transformada em uma variável aleatória normal padrão. Além disso, qualquer operação que possa ser aplicada a uma variável como esta irá se aplicar a qualquer outra normal após ser devidamente escalonada e deslocada.

Outras variáveis aleatórias contínuas

Existem muitas outras variáveis aleatórias contínuas e distribuições de probabilidade associadas que os ecólogos e estatísticos usam. Duas importantes são as variáveis aleatórias log-normais e variáveis aleatórias exponenciais (Figura 2.8). Uma **variável aleatória log-normal** X é uma variável aleatória tal como seu logaritmo natural $\ln(X)$ é uma variável aleatória normal. Muitas características ecológicas-chave de organismos, como massa corporal, são distribuídas de forma log-normal.[12]

Como a distribuição normal, a log-normal é descrita por dois parâmetros, μ e σ. O valor esperado de uma variável aleatória log-normal é:

$$E(X) = e^{\frac{2\mu+\sigma^2}{2}}$$

E a variância de uma distribuição log-normal é:

$$\sigma^2(X) = \left[e^{\frac{\mu+\sigma^2}{2}}\right]^2 \times \left[e^{\sigma^2} - 1\right]$$

Quando representados graficamente em escala logarítmica, a distribuição log-normal exibe uma curva em forma de sino. Contudo, se esses mesmos dados forem transformados de volta aos valores originais (aplicando a transformação e^X), a distribuição resultante é desviada, com uma longa cauda de probabilidade se estendendo à direita (*Ver* Capítulo 8 para mais exemplos em transformações de dados).

Variáveis aleatórias exponenciais são relacionadas às de Poisson. Relembre que estas descrevem o número de ocorrências raras (p. ex., contagens), como

[12] O exemplo ecológico mais familiar de uma distribuição log-normal é a da abundância relativa das espécies em uma comunidade. Se coletar uma amostra aleatória de indivíduos de uma comunidade, classificá-los de acordo com a espécie, e construir um histograma de frequência das espécies representadas em diferentes classes de abundância, os dados irão se assemelhar a uma distribuição normal quando as classes de abundância são representadas graficamente em escala logarítmica (Preston, 1948). Qual é a explicação para este padrão? Por um lado, muitos conjuntos de dados não biológicos (como a distribuição do patrimônio líquido entre países, ou a distribuição do tempo de duração de copos em um restaurante cheio) também seguem uma distribuição log-normal. Portanto, o padrão pode refletir uma resposta estatística genérica de populações aumentando exponencialmente (um fenômeno logarítmico) a muitos fatores independentes (May, 1975). Por outro lado, mecanismos biológicos específicos podem estar atuando, incluindo dinâmica de manchas (Ungland & Gray, 1982) ou partição hierárquica de nichos (Sugihara, 1980). O estudo de distribuições log-normal também é complicado por problemas amostrais. Como espécies raras em uma comunidade podem ser perdidas em amostras pequenas, a forma da distribuição de abundancia das espécies se altera com a intensidade da amostragem (Wilson, 1993). Não obstante, mesmo em comunidades bem amostradas, a cauda do histograma de frequências pode não se ajustar muito bem a uma distribuição log-normal verdadeira (Preston, 1981). Esta falta de ajuste pode surgir porque uma amostra grande de uma comunidade em geral conterá um misto de espécies residentes e em transição (Magurran & Henderson, 2003).

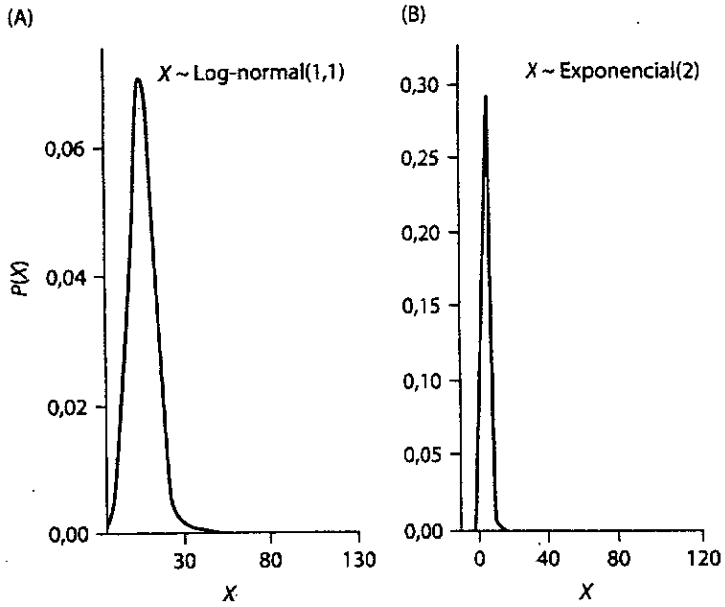


Figura 2.8 Distribuições log-normal e exponencial se ajustam a certos tipos de dados ecológicos, como a distribuição de abundâncias de espécies e distâncias de dispersão de sementes. (A) A distribuição log-normal é descrita por dois parâmetros, média e variância, ambos são 1 neste exemplo. (B) A distribuição exponencial é descrita por um único parâmetro *b*, que é 2 nesse exemplo. Ver a Tabela 2.4 para as equações usadas com as distribuições log-normal e exponencial. Ambas as distribuições, log-normal e exponencial, são assimétricas, com uma longa cauda a direita que desvia a distribuição à direita.

o número de visitas em um intervalo de tempo constante, ou o de indivíduos que ocorrem em uma área fixa. Os "espaços" entre variáveis aleatórias discretas, como o tempo ou a distância entre eventos de Poisson, podem ser descritos como variáveis aleatórias exponenciais contínuas. A função de distribuição de probabilidades de uma variável aleatória exponencial X tem apenas um parâmetro, geralmente escrita como β, e toma a forma $P(X) = \beta e^{-\beta X}$. O valor esperado de uma variável aleatória exponencial = $1/\beta$, e variância $1/\beta^2$.[13] A Tabela 2.4 traz um sumário das propriedades dessas distribuições contínuas comuns.

Outras variáveis aleatórias contínuas e suas funções de distribuição de probabilidades são usadas extensivamente em análises estatísticas. Elas incluem a distribuição-t de Student, chi-quadrado, F, gamma, gamma inversa e beta. Iremos discuti-las mais adiante, quando as usarmos em técnicas estatísticas específicas.

[13] É fácil simular uma variável aleatória contínua em seu computador, tirando vantagem do fato de que se U é uma variável aleatória uniforme definida no intervalo unitário fechado $\{0,1\}$, então $-\ln(U)/\beta$ é uma variável aleatória exponencial com parâmetro β.

TABELA 2.4 Quatro distribuições contínuas

Distribuição	Valor de probabilidade	$E(X)$	$\sigma^2(X)$	Comentários	Exemplo ecológico
Uniforme	$P(a<X<b)=1,0$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$	Use para resultados equiprováveis sobre o intervalo $[a,b]$.	Distribuição uniforme dos recursos.
Normal	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$	μ	σ^2	Gera uma curva simétrica em "forma de sino" para dados contínuos.	Distribuição do comprimento do espinho tibial e outras variáveis contínuas de tamanho.
Log-normal	$\frac{1}{\sigma X\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(X)-\mu}{\sigma}\right)^2}$	$e^{\frac{2\mu+\sigma^2}{2}}$	$\left[e^{\frac{\mu+\sigma^2}{2}}\right]^2\times\left[e^{\sigma^2}-1\right]$	Dados transformados em log de dados desviados para a direita frequentemente se ajustam à curva normal.	Distribuição das classes de abundância das espécies.
Exponencial	$P(X)=\beta e^{-\beta X}$	$1/\beta$	$1/\beta^2$	Distribuição contínua análoga à de Poisson.	Distância de dispersão de sementes.

A equação do valor de probabilidade determina a chance de obter um determinado valor X para cada distribuição. A expectativa $E(X)$ da distribuição de valores é estimada pela média ou média aritmética de uma amostra. A variância $\sigma^2(X)$ é uma medida da dispersão ou desvio das observações do $E(X)$. Essas distribuições são para variáveis em uma escala contínua que pode assumir qualquer número real, apesar de a distribuição exponencial ser limitada aos valores positivos.

O TEOREMA DO LIMITE CENTRAL

O **Teorema do Limite Central** é um dos pilares da probabilidade e análises estatísticas.[14] Segue uma breve descrição do teorema. Permita que S_n seja a soma ou a

Abraham De Moivre

Pierre Laplace

[14] A formulação inicial do Teorema do Limite Central foi contribuição de Abraham De Moivre (Informações biográficas na nota de rodapé 11) e Pierre Laplace. Em 1733, De Moivre provou sua versão do Teorema do Limite Central para uma variável aleatória de Bernoulli. Pierre Laplace (1749-1827) estendeu os resultados de De Moivre para qualquer variável aleatória binária. Laplace é mais lembrado por seu *Mécanique céleste* (Mecânica celeste), que traduziu o sistema geométrico dos estudos de mecânica de Newton para um sistema com base em cálculos. De acordo com Boyer, em *History of Mathematics* (1968), após Napoleão ter lido *Mécanique céleste* ele perguntou a Laplace por que não havia nenhuma menção a Deus no trabalho. Dizem que Laplace respondeu que ele não era necessário para aquela hipótese. Posteriormente, Napoleão apontou Laplace para ser Ministro do interior, mas mais tarde demitiu-o com o comentário de que "ele carregou o espírito do infinitamente pequeno para a gestão dos negócios" (*fide* Boyer, 1968). O matemático russo Pafnuty Chebyshev (1821-1884) provou o Teorema do Limite Central para qualquer variável aleatória, mas sua prova complexa é virtualmente desconhecida na atualidade. A prova moderna acessível do Teorema do Limite Central é contribuição de Andrei Markov (1856-1922) e Alexander Lyapounov (1857-1918), alunos de Chebyshev.

média aritmética de qualquer conjunto de n variáveis aleatórias X_i, independentes e distribuídas de forma idêntica:

$$S_n = \sum_{i=1}^{n} X_i$$

cada uma com o mesmo valor esperado μ e todas com a mesma variância σ^2. Então, S_n tem o valor esperado de $n\mu$ e variância $n\sigma^2$. Se padronizarmos S_n subtraindo o valor esperado de cada observação e dividindo pela raiz quadrada da variância,

$$S_{std} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

então, a distribuição de um conjunto de valores S_{pad} aproxima uma variável normal padrão.

Este é um resultado poderoso. O Teorema do Limite Central afirma que padronizar *qualquer* variável aleatória, que em si é a soma ou média aritmética de um conjunto de variáveis aleatórias independentes, resulta em uma nova variável, que é "quase o mesmo que"[15] uma normal padrão. Já utilizamos essa técnica quando geramos uma variável aleatória normal padrão Z a partir de uma variável aleatória normal (X). A beleza do Teorema do Limite Central é que ele nos permite usar ferramentas estatísticas que requerem que nossas observações amostrais sejam tiradas de um espaço amostral normalmente distribuído, mesmo que os dados subjacentes não o sejam. As únicas ressalvas são que o tamanho amostral precisa ser "suficientemente grande",[16] e que as observações devem ser independentes e tiradas a partir de uma distribuição com expectativa e variância comuns. Demonstraremos a importância do Teorema do Limite Central quando discutirmos as diferentes técnicas estatísticas usadas por ecólogos e cientistas da área ambiental (Capítulos 9-12).

[15] Mais formalmente, o Teorema do Limite Central declara que, para qualquer variável padronizada

$$Y_i = \frac{S_i - n\mu}{\sigma\sqrt{n}}$$

a área sob a distribuição de probabilidade normal padrão no intervalo (a,b) é igual a:

$$\lim_{i\to\infty} P(a < Y_i < b)$$

[16] Uma questão importante para ecólogos (e estatísticos) é com que velocidade a probabilidade $P(a < Y_i < b)$ converge para a área sob a distribuição de probabilidade normal padrão. Muitos ecólogos (e estatísticos) poderiam dizer que o tamanho amostral i deve ser no mínimo 10, mas estudos recentes sugerem que i deve exceder 10.000 antes de as duas convergirem até mesmo a duas casas decimais. Felizmente, a maioria dos testes estatísticos vai ao encontro da premissa de normalidade, de forma que podemos usar o Teorema do Limite Central mesmo que nossos dados padronizados não exibam com perfeição uma distribuição normal. Hoffman Jørgensen (1994; *ver também* o Capítulo 5) fez uma revisão minuciosa e técnica do Teorema do Limite Central.

RESUMO

Variáveis aleatórias assumem uma variedade de medidas, mas suas distribuições podem ser caracterizadas por sua expectativa e variância. Distribuições discretas, como as de Bernoulli, binomial e Poisson, se aplicam a dados que são contagens discretas, enquanto que distribuições contínuas, como a uniforme, normal e exponencial, se aplicam a dados medidos em uma escala contínua. Independente da distribuição subjacente, o Teorema do Limite Central afirma que as somas ou médias aritméticas de amostras grandes e independentes seguem a distribuição normal se forem padronizadas. Para uma ampla variedade de dados, incluindo aqueles coletados mais comumente por ecólogos e cientistas da área ambiental, o Teorema do Limite Central apoia o uso de testes estatísticos que assumem distribuição normal.

3

Estatísticas Descritivas: Medidas de Posição e Dispersão

Os dados são a essência das investigações científicas, porém raramente mostramos todos os que coletamos. Ao invés disso, resumimos nossos dados utilizando **estatísticas descritivas**. Biólogos e estatísticos distinguem entre duas formas de estatísticas descritivas: medidas de **posição** e medidas de **dispersão**. As medidas de posição ilustram onde a maioria dos dados se encontra; essas medidas incluem as médias, medianas e modas. Por outro lado, medidas de dispersão descrevem o quanto os dados são variáveis; essas medidas incluem o desvio-padrão amostral, a variância e o erro-padrão. Apresentaremos as estatísticas descritivas mais comuns e ilustraremos como elas surgem diretamente da **Lei dos Grandes Números**, um dos mais importantes teoremas da probabilidade.

De agora em diante, adotaremos a notação estatística padrão quando descrevermos variáveis aleatórias e quantidades estatísticas ou estimadores. Variáveis aleatórias serão designadas como Y, em que cada observação individual é indexada com um subscrito, Y_i. O subscrito i indica a i-ésima observação. O tamanho da amostra será denotado por n, de forma que o i pode assumir qualquer valor entre 1 e n. A média aritmética é escrita como $\bar{Y}$. **Parâmetros** desconhecidos (ou população estatística) de distribuições, como valores esperados e variâncias, serão escritos com letras gregas (como μ para o valor esperado, σ^2 para a variância esperada e σ para o desvio-padrão esperado), enquanto os estimadores estatísticos desses parâmetros (com base em dados reais) serão escritos com letras em itálico (como $\bar{Y}$ para a média aritmética, s^2 para a variância da amostra e s para o desvio-padrão amostral).

Neste Capítulo usaremos como exemplo os dados ilustrados na Figura 2.6, as medidas simuladas do comprimento do espinho tibial de 50 aranhas Linyphiidae. Esses dados, colocados em ordem ascendente, são ilustrados na Tabela 3.1.

TABELA 3.1 Medidas, em ordem crescente, dos espinhos tibiais de 50 aranhas Linyphiidae (em milímetros)

0,155	0,207	0,219	0,228	0,241	0,249	0,263	0,276	0,292	0,307
0,184	0,208	0,219	0,228	0,243	0,250	0,268	0,277	0,292	0,308
0,199	0,212	0,221	0,229	0,247	0,251	0,270	0,280	0,296	0,328
0,202	0,212	0,223	0,235	0,247	0,253	0,274	0,286	0,301	0,329
0,206	0,215	0,226	0,238	0,248	0,258	0,275	0,289	0,306	0,368

Este conjunto de dados simulados é utilizado ao longo deste Capítulo para ilustrar medidas de estatísticas descritivas e distribuição de probabilidades. Embora dados brutos deste tipo formem as bases de todos os cálculos na estatística, raramente são publicados, pois eles são muito extensos e difíceis de compreender. Estatísticas descritivas, se usadas de maneira adequada, comunicam e descrevem com concisão os padrões nos dados brutos sem enumerar cada observação individual.

MEDIDAS DE POSIÇÃO

A média aritmética

Existem diversas formas de descrever um conjunto de dados. A mais familiar é a **média aritmética** das observações, que é calculada pela soma das observações (Y_i), dividida pelo número de observações (n) e é denotada por $\bar{Y}$:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{3.1}$$

Para os dados da Tabela 3.1 $\bar{Y} = 0{,}253$. A Equação 3.1 parece similar, mas não é equivalente à Equação 2.6, que foi utilizada no Capítulo 2 para calcular o valor esperado de uma variável aleatória discreta:

$$E(Y) = \sum_{i=1}^{n} Y_i p_i$$

onde os Y_i são os valores que a variável aleatória pode ter, e os p_i são suas probabilidades. Para uma variável contínua na qual cada Y_i ocorre apenas uma vez, com $p_i = 1/n$, as Equações 3.1 e 2.6 dão resultados idênticos.

Por exemplo, deixe que *comprimento do espinho* seja um conjunto constituído pelas 50 observações da Tabela 3.1: *Comprimento do espinho* = {0,155, 0,184, ..., 0,329, 0,368}. Se cada elemento (ou evento) dentro do *comprimento do espinho* é independente dos outros, então a probabilidade p_i de qualquer uma dessas 50 observações independentes é 1/50. Usando a Equação 2.6, podemos calcular o valor esperado do *comprimento do espinho* como:

$$E(Y) = \sum_{i=1}^{n} Y_i p_i$$

onde Y_i é o i-ésimo elemento e $p_i = 1/50$. Esta soma:

$$E(Y) = \sum_{i=1}^{n} Y_i \times \frac{1}{50}$$

agora é equivalente à Equação 3.1, usada para calcular a média aritmética de n observações de uma variável aleatória Y.

$$\bar{Y} = \sum_{i=1}^{n} p_i Y_i = \sum_{i=1}^{n} Y_i \times \frac{1}{50} = \frac{1}{50} \sum_{i=1}^{n} Y_i$$

Para calcular o valor esperado de *comprimento do espinho,* usamos a fórmula para o valor esperado de uma variável aleatória discreta (Equação 2.6). Contudo, os dados da Tabela 3.1 representam observações de uma variável aleatória contínua normal. Tudo que sabemos sobre o valor esperado dessa variável é que ele tem algum valor verdadeiro subjacente, que denotamos como μ. O valor da média calculado para o *comprimento do espinho* tem alguma relação com o valor desconhecido de μ?

Se três condições forem satisfeitas, a média aritmética das observações em nossa amostra é um **estimador imparcial** de μ. Essas três condições são:

1. As observações são feitas em indivíduos escolhidos de maneira aleatória.
2. As observações na amostra são independentes umas das outras.
3. As observações são realizadas em uma população maior, que pode ser descrita por uma variável aleatória normal.

O fato de que a $\bar{Y}$ de uma amostra aproxima a μ da população, da qual foi retirada, é um caso especial do segundo teorema fundamental da probabilidade, a **Lei dos Grandes Números.**[1]

Eis uma descrição da Lei dos Grandes Números. Considere um conjunto infinito de amostras aleatórias de tamanho n, tiradas de uma variável aleatória Y. Assim, Y_1 é uma amostra de Y com um dado, $\{y_1\}$. Então, Y_2 é uma amostra de tamanho 2, $\{y_1, y_2\}$, etc. A Lei dos Grandes Números estabelece que, conforme o tamanho amostral n aumenta, a média aritmética de Y_i (Equação 3.1) aproxima-se do valor esperado de Y, $E(Y)$. Em notação matemática, escrevemos:

Andrei Kolmogorov

[1] A versão moderna (ou "poderosa") da Lei dos Grande Números foi provada pelo matemático russo Andrei Kolmogorov (1903-1987), que também estudou os **processos de Markov** como aqueles utilizados na moderna computação de análises bayesianas (*ver* Capítulo 5) e mecânica de fluidos.

$$\lim_{n\to\infty}\left(\frac{\sum_{i=1}^{n} y_i}{n} = \bar{Y}_n\right) = E(Y) \qquad (3.2)$$

Em palavras, dizemos que conforme n se torna muito grande, a média aritmética das Y_i é igual a $E(Y)$ (*ver* Figura 3.1).

Em nosso exemplo, os comprimentos do espinho tibial de todos os indivíduos de aranhas Linyphiidae em uma população podem ser descritos como uma variável aleatória normal com valor esperado = μ. Não podemos medir todos os (muitos) espinhos, mas sim um subconjunto deles; a Tabela 3.1 traz $n = 50$ dessas medidas. Se cada espinho foi medido em uma aranha individual, se cada aranha foi escolhida aleatoriamente para ser medida e se não há nenhum viés em nossas medidas, então o valor esperado de cada observação deve ser o mesmo (pois eles fazem parte da mesma população, infinitamente grande, de aranhas). A Lei dos Grandes Números diz que o comprimento médio dos espinhos de nossas 50 medidas se aproxima do valor esperado do comprimento dos espinhos em toda a população. Por isso, podemos estimar o valor esperado μ desconhecido com a média aritmética de nossas observações. Como mostra a Figura 3.1, a estimativa da verdadeira média da população torna-se mais confiável conforme acumulamos mais dados.

Outras médias

A média aritmética não é a única medida de posição de um conjunto de dados. Em alguns casos, irá produzir respostas inesperadas. Por exemplo, suponha que uma população de veado-mula* (*Odocoileus hemionus*) aumenta 10% em tamanho em um ano e 20% no próximo ano. Qual é a taxa de crescimento médio da população a cada ano?[2] A resposta não é 15%!

Você pode ver esta diferença trabalhando com alguns números. Suponha que o tamanho inicial da população é de 1.000 veados. Após um ano, seu ta-

[2] Nesta análise, usamos a taxa finita de crescimento, λ, como parâmetro do crescimento populacional. Trata-se de um multiplicador que opera no tamanho populacional a cada ano, como $N_{t+1} = \lambda N_t$. Assim, se a população cresce 10% a cada ano, λ = 1,10, ou se a população decresce 5% a cada ano, λ = 0,95. Uma medida similar à taxa de crescimento populacional é a taxa de crescimento instantâneo, r, na qual as unidades são indivíduos/(indivíduos × tempo). Matematicamente, $\lambda = e^r$ e $r = \ln(\lambda)$. *Ver* Gotelli (2001**) para mais detalhes.

* N. de T. O veado-mula (Odocoileus hemionus) é um veado norte-americano. Ele é encontrado em todo o oeste da América do Norte, desde o sul de Yukon, no Alasca, até a baixa califórnia e o norte do México.

** N. deT. Este livro já está na 4ª edição, publicada em 2008. A obra foi traduzida para o português pela editora PLANTA, com o título de *Ecologia*.

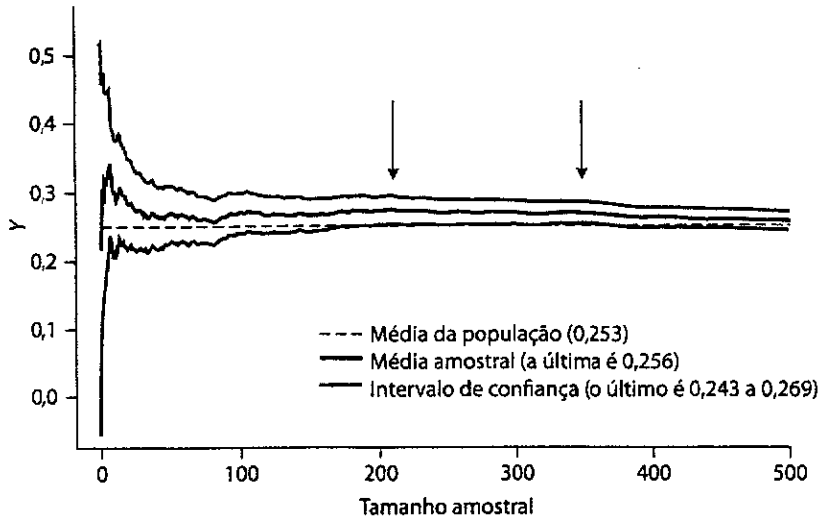


Figura 3.1 Ilustração da Lei dos Grandes Números e a construção de intervalos de confiança usando os dados dos espinhos tibiais de aranhas da Tabela 3.1. A média da população (0,253) é indicada pela linha pontilhada. A média amostral para amostras de tamanhos crescentes (n) é indicada pela linha sólida central e ilustra a Lei dos Grandes Números: conforme o tamanho amostral aumenta, a média amostral se aproxima da verdadeira média da população. As linhas sólidas superior e a inferior ilustram o intervalo de confiança de 95% ao redor da média. A largura do intervalo de confiança decresce conforme o tamanho amostral aumenta. Intervalos de confiança de 95% construídos dessa forma devem conter a verdadeira média da população. Note, contudo, que existem amostras (entre as setas) para as quais o intervalo de confiança não inclui a verdadeira média da população. As curvas foram construídas usando algoritmos e códigos do S-Plus publicados por Blume e Royal (2003).

manho (N_1) será de (1,10) × 1.000 = 1.100. Após o segundo ano, o tamanho populacional (N_2) será (1,20) × 1.100 = 1.320. Contudo, se a taxa de crescimento médio por ano for de 15%, o tamanho populacional será de (1,15) × 1.000 = 1.150 após um ano e (1,15) × 1.150 = 1.322,50 após dois anos. Esses números são próximos, mas não idênticos; após mais alguns anos, os resultados irão divergir substancialmente.

A MÉDIA GEOMÉTRICA No Capítulo 2, apresentamos a distribuição log-normal: se Y é uma variável aleatória com distribuição log-normal, então a variável aleatória $Z = \ln(Y)$ é uma variável aleatória normal. Se calcularmos a média aritmética de Z,

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \qquad (3.3)$$

qual é o valor expresso em unidades de Y? Primeiro, reconheça que se $Z = \ln(Y)$, então $Y = e^Z$, onde e é a base do logaritmo natural e é igual a ~2,71828.... Assim, o

valor de $\bar{Z}$ em unidades de Y é $e^{\bar{Z}}$. Essa **média de transformação reversa** é chamada de **média geométrica** e é escrita como MG_Y.

A forma mais simples para calcular a média geométrica é tirar o antilog da média aritmética:

$$MG_Y = e^{\left[\frac{1}{n}\sum_{i=1}^{n}\ln(Y_i)\right]} \tag{3.4}$$

Uma boa característica é que a soma dos logaritmos de um conjunto de números se iguala ao de seus produtos: $\ln(Y_1) + \ln(Y_2) + \ldots = \ln(Y_1 Y_2 \ldots Y_n)$. Portanto, outra forma de calcular a média geométrica é tirar a n-ésima raiz do produto das observações:

$$MG_Y = \sqrt[n]{Y_1 Y_2 \ldots Y_n} \tag{3.5}$$

Da mesma forma que temos um símbolo especial para adicionar uma série de números:

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + \ldots + Y_n$$

também temos um símbolo especial para a multiplicação de uma série de números:

$$\prod_{i=1}^{n} Y_i = Y_1 \times Y_2 \times \ldots \times Y_n$$

Assim, também podemos escrever a fórmula da média geométrica como:

$$MG_Y = \sqrt[n]{\prod_{i=1}^{n} Y_i}$$

Vamos ver se a média geométrica da taxa de crescimento populacional faz um trabalho melhor na predição da taxa de crescimento médio que a média aritmética faz. Primeiro, se expressarmos as taxas de crescimento populacional como multiplicadores, as taxas de crescimento anual de 10% e de 20% se tornam 1,10 e 1,20, e o logaritmo natural desses dois valores é $\ln(1{,}10) = 0{,}09531$ e $\ln(1{,}20) = 0{,}18232$. A média aritmética desses dois valores é 0,138815. O cálculo reverso nos dá a média geométrica de $MG_Y = e^{0{,}138815} = 1{,}14891$, que é um pouco menor que a média aritmética de 1,20.

Agora podemos calcular a taxa de crescimento populacional ao longo de dois anos, usando a taxa de crescimento da média geométrica. No primeiro ano, a população poderia crescer para $(1{,}14891) \times (1.000) = 1148{,}91$; no segundo ano, para $(1{,}14891) \times (1.148{,}91) = 1.319{,}99$. Esta é a mesma resposta que obtemos com um crescimento de 10% no primeiro ano e de 20% no segundo ano $[(1{,}10) \times (1.000) \times (1{,}20)] = 1.320$. Os valores iriam se igualar perfeitamente se as taxas de cresci-

mento não tivessem sido arredondadas. Note, também, que, apesar de o tamanho populacional ser sempre uma variável de números inteiros (0,91 veado pode ser visto somente em uma floresta teórica), nós a tratamos como uma variável contínua para ilustrar esses cálculos.

Por que a MG_Y nos dá a resposta correta? A razão é que o crescimento da população é um processo multiplicativo. Note que:

$$\frac{N_2}{N_0} = \left(\frac{N_2}{N_1}\right) \times \left(\frac{N_1}{N_0}\right) \neq \left(\frac{N_2}{N_1}\right) + \left(\frac{N_1}{N_0}\right)$$

Contudo, números que são multiplicados em uma escala aritmética podem ser adicionados em uma logarítmica. Assim

$$\ln\left[\left(\frac{N_2}{N_1}\right) \times \left(\frac{N_1}{N_0}\right)\right] = \ln\left(\frac{N_2}{N_1}\right) + \ln\left(\frac{N_1}{N_0}\right)$$

MÉDIA HARMÔNICA Um segundo tipo de média pode ser calculado de forma similar, usando a transformação do inverso ($1/Y$). O inverso da média aritmética dos inversos de um conjunto de observações é chamado de **média harmônica**:[3]

$$H_i = \frac{1}{\frac{1}{n}\sum \frac{1}{Y_i}} \qquad (3.6)$$

Para os dados dos espinhos da Tabela 3.1, $MG_Y = 0{,}249$ e $H_Y = 0{,}246$. Ambas as médias são menores que a aritmética (0,253); em geral, essas médias são ordenadas como $\bar{Y} > GM_Y > H_Y$. Contudo, se as observações são iguais ($Y_1 = Y_2 = Y_3 = \ldots Y_n$), as três médias são idênticas ($\bar{Y} = GM_Y = H_Y$).

[3] A média harmônica é usada em biologia da conservação e genética de populações para calcular o tamanho populacional efetivo, que é o equivalente a uma população com acasalamentos completamente ao acaso. Se o tamanho populacional efetivo é pequeno (< 50), mudanças ao acaso na frequência dos alelos, devido à deriva genética, são potencialmente importantes. Se o tamanho populacional muda de um ano para o outro, a média harmônica dá o tamanho populacional efetivo. Por exemplo, imagine que uma população estável de 100 lontras marinhas passa por um severo gargalo e é reduzida para um tamanho populacional de 12 em um único ano. Assim, os tamanhos da população são 100, 100, 12, 10, 100, 100, 100, 100, 100 e 100. A média aritmética é 91,2, mas a média harmônica é somente 57,6, sendo considerado o tamanho populacional efetivo no qual a deriva genética pode ser importante. A média harmônica não somente é menor que a aritmética, como também é especialmente sensível a valores extremos, que são pequenos. Nos séculos XVIII e XIX, as lontras marinhas da costa do Pacífico da América do Norte passaram por um severo gargalo populacional quando foram extensivamente caçadas. Apesar de as populações de lontras marinhas terem se recuperado em tamanho, ainda exibem pouca diversidade genética. (Larson et al., 2002.) (Fotografia de Warren Worthington, http://soundwaves.usgs.gov/2002/07/.)

Outras medidas de posição: a mediana e a moda

Ecólogos e cientistas da área ambiental em geral usam outras duas medidas de posição, a mediana e a moda, para descrever conjuntos de dados. A **mediana** é definida como o valor em um conjunto de observações ordenadas que tenha um número igual ao de observações acima e abaixo dele. Em outras palavras, a mediana divide um conjunto de dados em duas metades com o mesmo número de observações em cada uma. Para um número ímpar de observações, a mediana é apenas o valor central. Assim, se considerássemos apenas as 49 primeiras observações dos dados de comprimento do espinho, a mediana seria a 25ª observação (0,248). Porém, com um número par de observações, a mediana é definida como o ponto central entre (n/2)-ésima e [(n/2)+1]-ésima observação. Se considerássemos todas as 50 observações da Tabela 3.1, a mediana seria a média entre a 25ª e a 26ª observações, ou 0,2485.

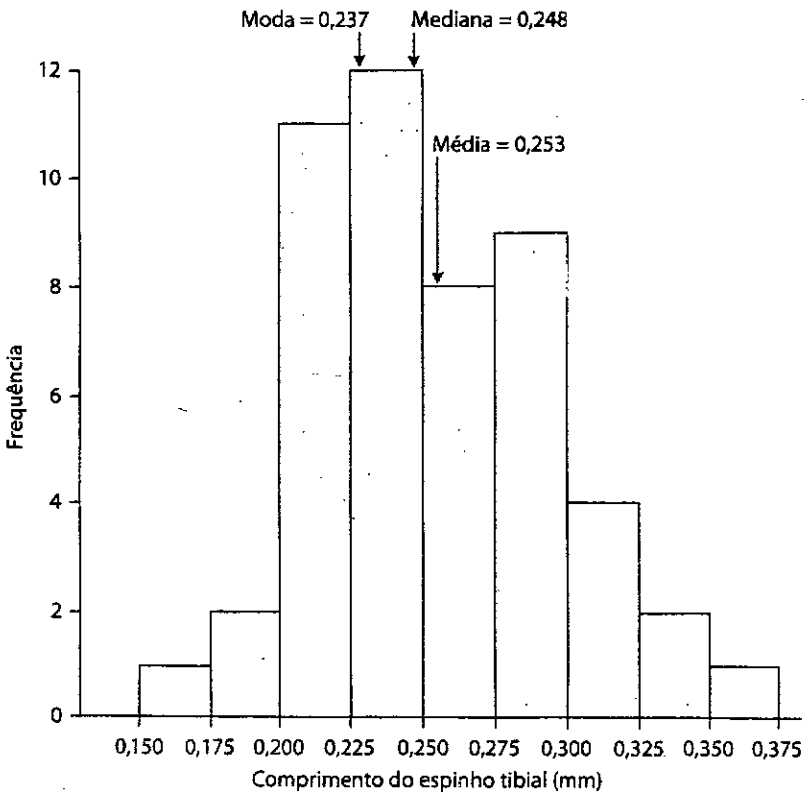


Figura 3.2 Histograma dos dados dos espinhos tibiais da Tabela 3.1 (n = 50) ilustrando a média aritmética, a mediana e a moda. A média é a expectativa dos dados, calculada como a média aritmética das medidas contínuas. A mediana é o ponto central do conjunto de observações ordenado. Metade de todas as observações é maior que a mediana e a outra metade é menor. A moda é a observação mais frequente.

A **moda**, por outro lado, é o valor das observações que ocorre mais frequentemente na amostra. A moda pode ser vista com facilidade em um histograma dos dados, já que ela é o pico. A Figura 3.2 ilustra a média aritmética, a mediana e a moda em um histograma com os dados dos espinhos tibiais.

Quando usar cada medida de posição

Por que escolher uma medida de posição em vez de outra? A média aritmética (Equação 3.1) é a medida de posição mais utilizada, em parte por ser familiar. A justificativa mais importante é que o Teorema do Limite Central (Capítulo 2) mostra que as médias aritméticas de amostras grandes provenientes de variáveis aleatórias obedecem a uma distribuição normal, ou gaussiana, mesmo que a variável aleatória subjacente não obedeça. Essa propriedade torna fácil testar hipóteses sobre médias aritméticas.

A média geométrica (Equações 3.4 e 3.5) é mais apropriada para descrever processos multiplicativos, como as taxas de crescimento populacional ou classes de abundância de espécies (antes elas são transformadas logaritmicamente; *ver* discussão da distribuição log-normal no Capítulo 2 e discussão de transformações de dados no Capítulo 8). A média harmônica (Equação 3.6) aparece em cálculos usados por geneticistas de populações e biólogos da conservação.

A mediana ou a moda descrevem melhor a posição dos dados quando a distribuição das observações não se ajusta a uma distribuição de probabilidades padrão, ou quando existem observações extremas. Isso acontece porque as médias aritmética, geométrica e harmônica são muito sensíveis a observações extremas (grandes ou pequenas), enquanto a mediana e a moda tendem a cair no meio da distribuição, independente da sua dispersão e forma. Em distribuições simétricas, como a distribuição normal, a média aritmética, a mediana e a moda são a mesma. Porém, em distribuições assimétricas, como a da Figura 3.2 (uma amostra aleatória relativamente pequena de uma distribuição normal subjacente), a média ocorre em direção à maior cauda da distribuição; a moda, na parte mais pesada da distribuição e a mediana, entre ambas.[4]

MEDIDAS DE DISPERSÃO

Não basta declarar a média ou outra medida de posição. Como há variação na natureza, e porque há um limite para a precisão com que conseguimos fazer me-

[4] As pessoas também usam diferentes medidas de posição para defender diferentes pontos de vista. Por exemplo, a renda familiar média nos Estados Unidos é consideravelmente maior que a típica (ou mediana). Isto porque a renda tem uma distribuição log-normal, de modo que as médias são ponderadas pela longa cauda à direita da curva, representando os muito ricos. Tenha muita atenção quando a média, a mediana ou a moda de um conjunto de dados é reportada, e desconfie se ela for reportada sem uma medida de dispersão ou variação.

didas, também é preciso quantificar e publicar a dispersão, ou variabilidade, de nossas observações.

A variância e o desvio-padrão

O conceito de variância foi apresentado no Capítulo 2. Para uma variável aleatória Y, a variância $\sigma^2(Y)$ é uma medida de quanto as observações dessa variável diferem do valor esperado. A variância é definida como $E[Y - E(Y)]^2$, onde $E(Y)$ é o valor esperado de Y. Com a média, a variância real de uma população é uma quantidade desconhecida. Da mesma forma que calculamos a $\bar{Y}$ estimada da média da população μ usando nossos dados, podemos calcular a s^2 estimada da variância da população σ^2 usando:

$$s^2 = \frac{1}{n}\sum(Y_i - \bar{Y})^2 \qquad (3.7)$$

Este valor também é chamado de **média quadrada**. Tal termo e seu acompanhante, a **soma dos quadrados**,

$$SQ_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad (3.8)$$

vão surgir novamente quando discutirmos regressão e análise de variância nos Capítulos 9 e 10. E tal como definimos o desvio-padrão σ de uma variável aleatória, como a raiz quadrada (positiva) de sua variância, podemos estimá-lo como $s = \sqrt{s^2}$. A transformação pela raiz quadrada garante que as unidades do desvio-padrão são as mesmas que as da média.

Notamos anteriormente que a média aritmética $\bar{Y}$ fornece uma estimativa imparcial da μ. Com isso queremos dizer que se amostrarmos a população repetidamente (infinitas vezes) e calcularmos a média aritmética (independente do tamanho amostral), a grande média desse conjunto de médias aritméticas deve se igualar a μ. Contudo, nossas estimativas iniciais da variância e do desvio-padrão não são estimadores imparciais da σ^2 e do σ, respectivamente. Em particular, a Equação 3.7 subestima a variância real da população.

O viés da Equação 3.7 pode ser ilustrado com um simples experimento imaginário. Suponha que você tire uma única observação Y_1 de uma população e tente estimar a μ e a $\sigma^2(Y)$. A sua estimativa da μ é a média de suas observações, que, neste caso, é apenas Y_1. Porém, se você estima a $\sigma^2(Y)$ usando a Equação 3.7, sua resposta sempre será 0,0, pois sua única observação é a mesma que sua estimativa da média. O problema é que, com um tamanho amostral de 1, já usamos nosso dado para estimar μ, e efetivamente não temos nenhuma informação adicional para estimar $\sigma^2(Y)$. Isso leva diretamente ao conceito de **graus de liberdade**, ou

seja, o número de peças de informações independentes que temos em um conjunto de dados para estimar parâmetros estatísticos. Em um conjunto de dados de tamanho amostral 1, não temos observações independentes o suficiente que possam ser utilizadas para estimar a variância.

A estimativa imparcial da variância, tratada como **variância amostral**, é calculada dividindo a soma dos quadrados por $(n - 1)$ em vez de dividir por n. Assim, a estimativa imparcial da variância é:

$$s^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2 \qquad (3.9)$$

e a estimativa imparcial do desvio-padrão, tratada como **desvio-padrão amostral**,[5] é:

$$s = \sqrt{\frac{1}{n-1}\sum(Y_i - \bar{Y})^2} \qquad (3.10)$$

As Equações 3.9 e 3.10 ajustam os graus de liberdade para o cálculo da variância amostral e do desvio-padrão. As equações também ilustram a necessidade de realizar pelo menos duas observações para estimar a variância de uma distribuição.

Para os dados dos espinhos tibiais da Tabela 3.1, $s^2 = 0{,}0017$ e $s = 0{,}0417$.

O erro-padrão da média

Outra medida de dispersão, usada frequentemente por ecólogos e cientistas da área ambiental, é o **erro-padrão da média**, a qual é abreviada como $s_{\bar{Y}}$ e é calculada pela divisão do desvio-padrão amostral pela raiz quadrada do tamanho amostral:

$$s_{\bar{Y}} = \frac{s}{\sqrt{n}} \qquad (3.11)$$

A Lei dos Grandes Números prova que, para um número infinitamente grande de observações, $\sum Y_i/n$ é uma aproximação da média da população μ, onde $Y_n = \{Y_i\}$ é uma amostra de tamanho n de uma variável aleatória Y com valor esperado

[5] O estimador imparcial do desvio-padrão é, por si só, imparcial apenas para amostras muito grandes ($n > 30$). Para tamanhos amostrais menores, a Equação 3.10 tende a subestimar o valor do σ da população (Gurland e Tripathi, 1971). Rohlf e Sokal (1995) fornecem uma tabela de consulta com o fator de correção pelo qual s deve ser multiplicado caso $n < 30$. Na prática, a maioria dos biólogos não aplica essas correções. Se o tamanho amostral dos grupos que estão sendo comparados não é muito diferente, nenhum problema grave acontece ao ignorar a correção do s.

$E(Y)$. De modo similar, a variância de $Y_n = \sigma^2/n$. Como o desvio-padrão é simplesmente a raiz quadrada da variância, o de Y_n é:

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

que é o mesmo que o erro-padrão da média. Portanto, temos uma estimativa do desvio-padrão da média da população μ, o erro-padrão.

Infelizmente, alguns cientistas não entendem a distinção entre o desvio-padrão (abreviado em legendas de figuras como DP*) e o erro-padrão da média (abreviado como EP*).[6] Como o erro-padrão da média sempre é menor que

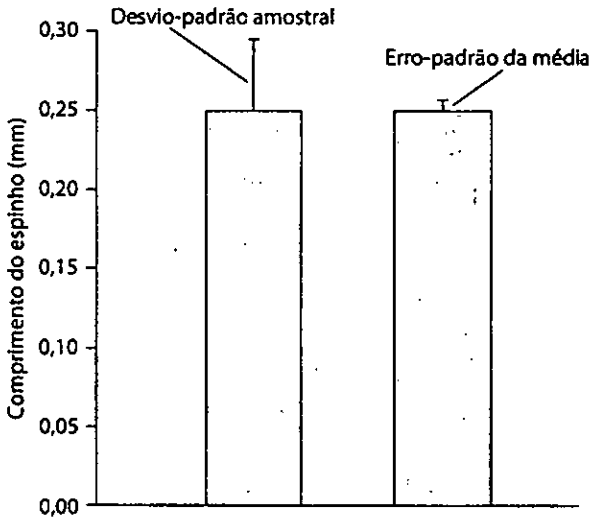


Figura 3.3 Diagrama de barras mostrando a média aritmética para os dados dos espinhos da Tabela 3.1 (n = 50), com barra de erro indicando o desvio-padrão amostral (barra esquerda) e o erro-padrão da média (barra direita). Enquanto o desvio-padrão mede a variabilidade de medidas individuais acerca da média, o erro-padrão mede a variabilidade da estimativa da média por si só. O erro-padrão se iguala ao desvio-padrão dividido por $\sqrt{n}$, então ele sempre será menor que o desvio-padrão, muitas vezes consideravelmente. As legendas e os textos de figuras devem sempre fornecer o tamanho amostral e indicar claramente quando o desvio-padrão ou o erro-padrão foi usado para construir as barras-de erros.

[6] Você deve ter notado que nos referimos ao erro-padrão da média e não simplesmente ao erro-padrão. O erro-padrão da média é o mesmo que o desvio-padrão de um conjunto de médias. Similarmente, podemos calcular o desvio-padrão de um conjunto de variâncias ou de outra estatística descritiva. Apesar de ser incomum ver outros erros-padrão em publicações da ecologia e ciências ambientais, pode haver casos em que será necessário reportar, ou no mínimo considerar, outros erros-padrão. Na Figura 3.1, o erro-padrão da mediana = 1,2533 × $s_{\bar{y}}$, e o erro-padrão do desvio-padrão = 0,7071 × $s_{\bar{y}}$. Sokal e Rohlf (1995) fornecem fórmulas para o erro-padrão de outras estatísticas comuns.

* N. de T. Em inglês *SD*, para abreviar "*Standard Deviation*"; SE, para abreviar "*Standard Error*".

o desvio-padrão amostral, as médias reportadas com o erro-padrão parecem menos variáveis que as reportadas com o desvio-padrão (Figura 3.3). Contudo, a decisão de quando apresentar o desvio-padrão amostral s ou o erro-padrão da média $s_{\bar{Y}}$ depende de qual inferência você quer que o leitor faça. Se suas conclusões, com base em uma única amostra, são representativas da população inteira, reporte o erro-padrão da média. Por outro lado, se as conclusões são limitadas às amostras que temos, é melhor reportar o desvio-padrão amostral. Pesquisas observacionais amplas cobrindo grandes escalas espaciais com um número significativo de amostras provavelmente são representativas da população de interesse como um todo (por consequência, reporte o $s_{\bar{Y}}$), enquanto experimentos pequenos e controlados, com poucas réplicas, provavelmente são baseados em um grupo único (e possivelmente pouco representativo) de indivíduos (por consequência, reporte o s).

Defendemos a idéia de reportar o desvio-padrão, s, que reflete com mais acurácia a variabilidade subjacente dos dados reais e dá menos crédito à generalidade. No entanto, ao passo que você fornece o tamanho amostral no texto, na figura ou na legenda da figura, os leitores podem calcular o erro-padrão da média a partir do desvio-padrão e vice-versa.

Assimetria (*skewness*), curtose e momento central

O desvio-padrão e a variância são casos especiais daquilo que os estatísticos (e físicos) chamam de **momento central** (MC). O MC é a média do desvio de todas as observações em um conjunto de dados da média das observações, elevado à potência r:

$$MC = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^r \qquad (3.12)$$

Na Equação 3.12, n é o número de observações, Y_i é o valor de cada observação individual, $\bar{Y}$ é a média aritmética das n observações e r é um inteiro positivo. O primeiro momento central ($r = 1$) é a soma das diferenças entre cada observação menos a média amostral (aritmética), que sempre é igual a 0. O segundo momento central ($r = 2$) é a variância (Equação 3.5).

O terceiro momento central ($r = 3$), dividido pelo desvio-padrão elevado ao cubo (s^3), é chamado de **assimetria** (denotada como g_1):

$$g_1 = \frac{1}{ns^3}\sum_{i=1}^{n}(Y_i - \bar{Y})^3 \qquad (3.13)$$

A assimetria descreve como a amostra difere em forma de uma distribuição simétrica. Uma distribuição normal tem $g_1 = 0$. Uma distribuição em que $g_1 > 0$ tem **assimetria à direita**: há uma longa cauda de observações maiores (i. e., à direita da) que a média. Em contraste, $g_1 < 0$ tem **assimetria à esquerda**:

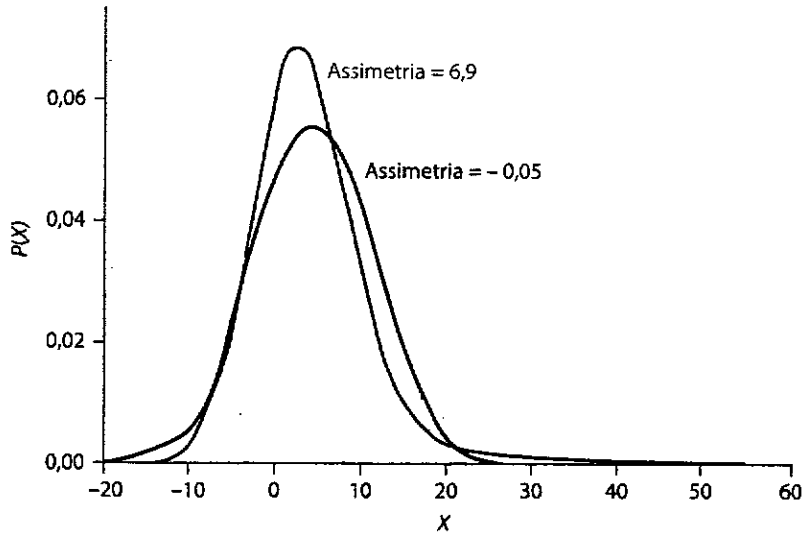


Figura 3.4 Distribuição contínua ilustrando a assimetria (g_1). A assimetria mede a extensão na qual a distribuição é assimétrica, com uma longa cauda de probabilidades à direita ou à esquerda. A curva verde é a distribuição log-normal ilustrada na Figura 2.8; ela tem assimetria positiva, com muito mais observações à direita da média do que à esquerda (uma longa cauda à direita), e uma medida de assimetria de 6,9. A curva preta representa uma amostra de 1.000 observações de uma variável aleatória normal com média e desvio-padrão idênticos aos da distribuição log-normal. Como esses dados foram tirados de distribuições normais simétricas, eles têm aproximadamente o mesmo número de observações em cada lado da média e a assimetria medida é aproximadamente 0.

há uma longa cauda de observações menores (i. e., à esquerda da) que a média (Figura 3.4).

A **curtose** tem sua base no quarto momento central ($r = 4$):

$$g_2 = \left[\frac{1}{ns^4}\sum_{i=1}^{n}(Y_i - \overline{Y})^4\right] - 3 \qquad \textbf{(3.14)}$$

A curtose mede a extensão na qual a densidade de probabilidade é distribuída nas caudas *versus* o centro da distribuição. Distribuições agregadas ou **platicúrticas** têm $g_2 < 0$; comparado à distribuição normal, existe mais massa de probabilidade no centro da distribuição e menos probabilidade nas caudas. Em contraste, distribuições **leptocúrticas** têm $g_2 > 0$. As distribuições leptocúrticas têm menos massa de probabilidade no centro e caudas de probabilidades relativamente pesadas (Figura 3.5).

Apesar da assimetria e da curtose terem sido reportadas com frequência na literatura ecológica antes da metade dos anos 1980, é incomum, na atualidade, ver

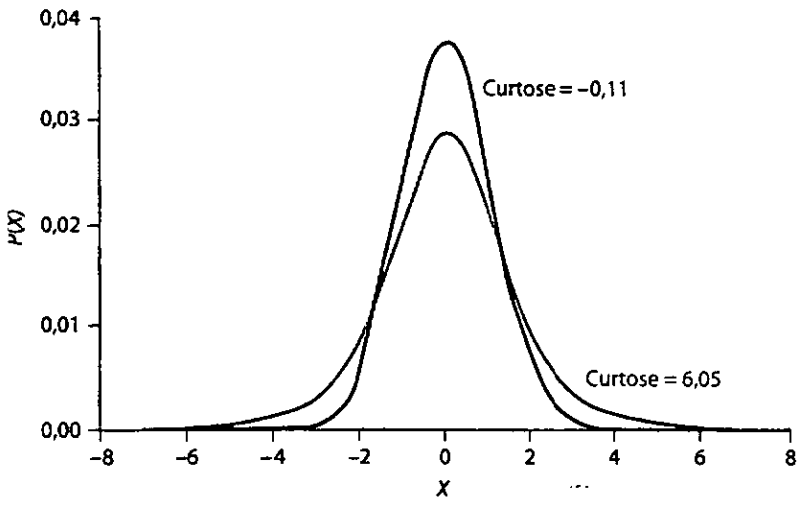


Figura 3.5 Distribuições ilustrando curtoses (g_2). A curtose mede a extensão na qual a distribuição é de cauda-pesada ou de cauda-leve, quando comparadas a uma distribuição normal padrão. Distribuições com caudas-pesadas são leptocúrticas e contêm relativamente mais área nas caudas da distribuição e menos no centro. Distribuições leptocúrticas possuem valores positivos para g_2. Distribuições com caudas-leves são platicúrticas e contêm relativamente menos área nas caudas da distribuição e mais no centro. Distribuições platicúrticas têm valores negativos para g_2. A curva preta representa uma amostra com 1.000 observações de uma variável aleatória normal com média 0 e desvio-padrão 1 ($X \sim N(0,1)$); sua curtose é próxima de 0. A curva verde é uma amostra de 1.000 observações de uma distribuição t com 3 graus de liberdade. A distribuição t é leptocúrtica e tem curtose positiva ($g_2 = 6,05$ neste exemplo).

esses valores reportados. Suas propriedades estatísticas não são boas: são muito sensíveis a medidas discrepantes (*outliers*) e a diferenças na média da distribuição. Weiner e Solbrig (1984) discutem o problema de usar medidas de assimetria em estudos ecológicos.

Quantis

Outra forma de ilustrar a dispersão de uma distribuição é reportar os **quantis**. Estamos todos familiarizados com um tipo de quantil, o **percentil**, por causa de seu uso em testes padronizados. Quando o escore de um teste é reportado dentro do 90° percentil, 90% dos escores são menores que aquele que foi reportado, e 10% são maiores. Neste capítulo vimos outro exemplo de um percentil – a mediana, que é o valor localizado no 50° percentil dos dados. Em apresentações de dados estatísticos reportamos mais comumente o **quartil** superior e o inferior – os valores do 25° e do 75° percentil, e o **decil** superior e o inferior – os valores do 10° e do 90° percentil. Esses valores para os dados dos espinhos são ilustrados, de forma concisa, em um ***box plot*** (Figura 3.6). Ao contrário da variância e do desvio-

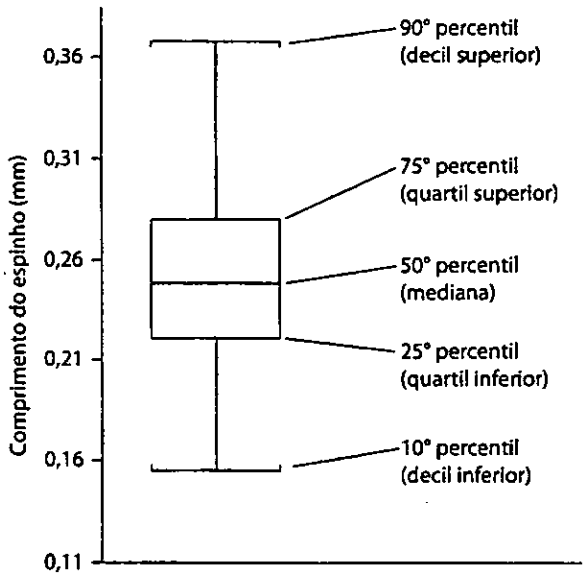


Figura 3.6 *Box plot* ilustrando os quantis dos dados da Tabela 3.1 ($n = 50$). A linha indica o 50° percentil (mediana) e a "caixa" engloba 50% dos dados, a partir do 25° até o 75° percentil. As linhas verticais se estendem do 10° ao 90° percentil.

-padrão, os valores dos quantis não dependem dos valores da média aritmética ou da mediana. Quando as distribuições são assimétricas ou possuem **medidas discrepantes – *outliers*** (dados com valores extremos que não são característicos da distribuição da qual foram amostrados, *ver* Capítulo 8), os *box plots* dos quantis podem retratar a distribuição dos dados de forma mais acurada que gráficos convencionais de média e desvio-padrão.

Usando as medidas de dispersão

Por si só, as medidas de dispersão não são especialmente informativas. A sua utilidade primária é para comparar dados de diferentes populações aos de diferentes tratamentos dentro de experimentos. Por exemplo, a análise de variância (Capítulo 10) usa os valores da variância amostral para testar hipóteses de que os tratamentos experimentais diferem uns dos outros. O conhecido teste-*t* usa o desvio-padrão amostral para testar hipóteses de que as médias de duas populações diferem uma da outra. Não é simples comparar apenas a variabilidade sobre populações ou grupos de tratamentos, porque a variância e o desvio-padrão dependem da média amostral. Contudo, descontando o desvio-padrão pela média, podemos calcular uma medida independente de variabilidade, chamada de **coeficiente de variação**, ou *CV*.

O *CV* é o desvio-padrão amostral dividido pela média, $s/\bar{Y}$, e é convencionalmente multiplicado por 100 para dar a porcentagem. O *CV* para os dados dos espinhos = 16,5%. Se outra população de aranhas tinha um *CV* do comprimento

do espinho tibial = 25%, poderíamos dizer que a primeira população é menos variável que a segunda população.

Um índice parecido é o **coeficiente de dispersão**, que é calculado através da divisão da variância amostral pela média ($s^2/\bar{Y}$). O coeficiente de dispersão pode ser usado com dados discretos para avaliar quando os indivíduos estão agregados ou hiper-dispersos no espaço, ou quando estão dispersos de maneira aleatória, como predito pela distribuição de Poisson. Forças biológicas que violam a independência farão com que as distribuições observadas sejam diferentes daquelas preditas pela de Poisson.

Por exemplo, algumas larvas de invertebrados marinhos exibem uma resposta à fixação agregada: uma vez que um juvenil ocupa uma mancha, ela se torna muito atrativa enquanto superfície de fixação para as larvas subsequentes (Crisp, 1979). Comparada a uma distribuição de Poisson, essas distribuições agregadas ou amontoadas tenderão a ter muitas amostras com alto número de ocorrência e muitas outras com zero ocorrência. Em contraste, muitas colônias de formigas exibem forte territorialidade e matam ou expulsam outras que tentem estabelecer colônias dentro do território (Levings e Traniello, 1981). Esse comportamento de segregação também puxa a distribuição para fora da de Poisson. Neste caso, as colônias serão hiperdispersas: haverá poucas amostras na classe de freqüência 0 e outras poucas com alto número de ocorrências.

Como a variância e a média de uma variável aleatória de Poisson se igualam a λ, o coeficiente de dispersão (CD) para uma variável aleatória de Poisson = $\lambda/\lambda = 1$. Por outro lado, se os dados são amontoados ou agregados, $CD > 1{,}0$, e se os dados são hiperdispersos ou segregados, $CD < 1{,}0$. Contudo, a análise do padrão espacial com distribuição de Poisson pode se tornar complicada, porque os resultados dependem não somente do grau de agregação, ou segregação dos organismos, como também do tamanho, do número e da colocação das unidades amostrais. Hulbert (1990) discute alguns dos temas envolvidos em ajustes de dados espaciais à distribuição de Poisson.

ALGUMAS QUESTÕES FILOSÓFICAS EM TORNO DAS ESTATÍSTICAS DESCRITIVAS

As estatísticas descritivas fundamentais – a média amostral, o desvio-padrão e a variância – são estimativas dos **parâmetros** reais no nível de população, μ, σ e σ^2 que obtemos diretamente a partir dos dados. Como é impossível amostrar a população inteira, somos forçados a estimar esses parâmetros desconhecidos pela $\bar{Y}, s, s^2$. Fazendo isso, atingimos um pressuposto fundamental: que existe um valor fixo e verdadeiro para cada um desses parâmetros. A Lei dos Grandes Números prova que, se amostrarmos uma população infinitas vezes, a média dos muitos $\bar{Y}$'s que calcularmos a partir de nossas amostras se igualaria a μ.

A Lei dos Grandes Números forma os alicerces daquilo que passou a ser conhecido como **estatística paramétrica, frequentista, ou assintótica**. A estatística paramétrica é assim chamada por causa do pressuposto de que a variável medida pode ser descrita por uma aleatória ou uma distribuição de

probabilidade, de forma conhecida e com parâmetros fixos definidos. As estatísticas, frequentista e assintótica, são assim chamadas porque assumimos que se o experimento for repetido muitas vezes, as estimativas mais frequentes dos parâmetros poderiam convergir em (atingir uma assíntota em) seus valores verdadeiros.

Mas e se esse pressuposto fundamental – de que os parâmetros subjacentes possuem valores verdadeiros e fixos – for falso? Por exemplo, se nossas amostras foram realizadas em um período longo de tempo, poderia haver mudanças no comprimento dos espinhos das tíbias das aranhas por causa de plasticidade fenotípica no crescimento, ou mesmo mudança evolucionária devido à seleção natural. Ou, talvez nossas amostras tenham sido realizadas em um período curto de tempo, mas cada aranha tenha vindo de um micro-hábitat distinto, para os quais existe uma única expectativa e variância para comprimento da tíbia. Nesse caso, existe algum significado real para um único valor da estimativa do comprimento médio de um espinho tibial? A estatística bayesiana parte do pressuposto fundamental de que parâmetros no nível de população, como μ, σ, σ^2, são, por si só, variáveis aleatórias. Uma análise bayesiana produz estimativas não apenas dos valores dos parâmetros, mas também da variabilidade inerente desses parâmetros.

A distinção entre as abordagens frequentista e bayesiana está longe de ser trivial, e resultou em muitos anos de debates acrimoniosos, primeiro entre estatísticos e mais recentemente entre ecólogos. Como veremos no Capítulo 5, as estimativas bayesianas de parâmetros como variáveis aleatórias frequentemente requerem complexos cálculos de computador. Em contraste, estimativas frequentistas de parâmetros, como valores fixos, usam fórmulas simples, as quais esboçamos neste capítulo. Devido à complexidade computacional das estimativas bayesianas, no início era difícil saber quando os resultados de análises frequentistas e bayesianas seriam quantitativamente diferentes. Contudo, o avanço tecnológico permitiu executar análises bayesianas complexas. Sob certas condições, o resultado dos dois tipos de análises é quantitativamente similar. A decisão de qual tipo utilizar, portanto, deve basear-se mais por um ponto de vista filosófico que pelo resultado quantitativo (Ellison, 2004). Entretanto, a interpretação dos resultados estatísticos pode ser completamente diferente entre ambas as perspectivas. Um exemplo dessa diferença é a construção e interpretação dos intervalos de confiança para a estimativa de parâmetros.

INTERVALOS DE CONFIANÇA

Os cientistas frequentemente usam o desvio-padrão amostral para construir um **intervalo de confiança** ao redor da média (Figura 3.1). Para uma variável aleatória que segue a distribuição normal, aproximadamente 67% das observações ocorrem dentro de ± 1 desvios-padrão da média, e, em média, 96% ocorrem den-

tro de ± 2 desvios-padrão da média.[7] Usamos essa observação para criar um intervalo de confiança de 95%, o que para amostras grandes é o intervalo delimitado por $(\overline{Y} - 1{,}96s_{\overline{Y}}, \overline{Y} + 1{,}96s_{\overline{Y}})$. O que este intervalo representa? Ele significa que a probabilidade de que a verdadeira média da população μ caia dentro do intervalo de confiança = 0,95.

$$P(\overline{Y} - 1{,}96s_{\overline{Y}} \leq \mu \leq \overline{Y} + 1{,}96s_{\overline{Y}}) = 0{,}95 \qquad (3.15)$$

Como a nossa média amostral e o desvio-padrão amostral da média são derivados de um único exemplar, o intervalo de confiança mudará se testarmos a população novamente (mas se o método é aleatório e sem viés, não deveria mudar muito). Assim, essa expressão está afirmando que a probabilidade de que a verdadeira média da população μ caia dentro de um único intervalo de confiança calculado = 0,95. Consequentemente, se formos amostrar repetidas vezes a população (mantendo o tamanho amostral constante), 5% das vezes poderíamos esperar que a verdadeira média μ ficasse fora do intervalo de confiança.

Interpretar um intervalo de confiança é complicado. Uma interpretação errônea é "Existe 95% de chance de que a verdadeira média da população μ ocorra dentro desse intervalo". O intervalo de confiança contém ou não contém μ; diferente do gato quântico de Schrödinger (*ver* Nota de rodapé 9, Capítulo 1), μ não pode estar simultaneamente dentro e fora do intervalo de confiança. O que você pode dizer é que, em 95% das vezes, um intervalo calculado desta forma irá conter o valor fixo de μ. Desta forma, se você realizar seu experimento 100 vezes, e criar 100 intervalos de confiança, aproximadamente 95 deles irão conter μ e 5 não (*ver* Figura 3.1 para um exemplo de quando um intervalo de confiança de 95% não inclui a verdadeira média da população). Blume e Royal (2003) fornecem exemplos adicionais e uma descrição pedagógica mais detalhada.

Esta explicação bastante complicada não é satisfatória, e não é exatamente o que você gostaria de afirmar quando construir um intervalo de confiança. Intuitivamente, você diria o quanto está confiante de que a média está dentro do intervalo (i. e., você está 95% certo de que a média está no intervalo). Um estatístico

[7] Use a "regra dos dois desvios-padrão" quando ler a literatura científica e tome por hábito rapidamente estimar intervalos de confiança grosseiros para dados amostrais. Por exemplo, suponha que você lê em um artigo que a média do conteúdo de nitrogênio de uma amostra de tecidos de plantas foi de 3,4% ± 0,2, sendo 0,2 é o desvio-padrão amostral. Dois desvios-padrão = 0,4, valor que é então somado e subtraído da média. Portanto, aproximadamente 95% das observações foram entre 3,0% e 3,8%. Você pode usar essa mesma dica quando examinar gráficos de barras, nos quais o desvio-padrão é representado como barra de erro. Eis uma excelente forma de usar estatísticas descritivas para verificar no ato as diferenças estatísticas reportadas entre grupos.

frequentista, contudo, não pode afirmar isso. Se existe uma média μ fixa da população, ou ela está ou não está dentro do intervalo e a declaração da probabilidade (Equação 3.15) afirma a probabilidade desse intervalo de confiança de incluir μ. Por outro lado, um estatístico bayesiano se refere a **intervalos de credibilidade**, para distingui-lo do intervalo de confiança frequentista. *Ver* Capítulo 5 e Ellison (1996) para detalhes adicionais.

Intervalos de confiança generalizados

Podemos construir qualquer intervalo de confiança com base nos percentis, como de 90%, ou de 50%. A fórmula geral para um n% intervalo de confiança é:

$$P(\bar{Y} - t_{\alpha[n-1]} s_{\bar{Y}} \leq \mu \leq \bar{Y} + t_{\alpha[n-1]} s_{\bar{Y}}) = (1-\alpha) \qquad (3.16)$$

onde $t_{\alpha[n-1]}$ é o valor crítico de uma **distribuição-*t*** com probabilidade $P = \alpha$, e tamanho amostral n. Essa probabilidade expressa a percentagem de área sob as duas caudas da curva de uma distribuição-*t* (Figura 3.7). Para uma curva normal padrão (uma distribuição-*t* com $n = \infty$), 95% da área sob a curva está dentro de ±1,96 desvio-padrão da média. Deste modo, 5% da área ($P = 0,05$) sob a curva permanecem nas duas caudas para além dos pontos a ± 1,96.

Então o que é a distribuição-*t*? Relembre do Capítulo 2 que uma transformação aritmética de uma variável aleatória normal é, por si só, uma variável aleatória normal. Considere o conjunto de médias amostrais $\{\bar{Y}_k\}$ resultante de um conjunto de grupos de medidas replicadas de uma variável aleatória normal com média μ desconhecida. O desvio das médias amostrais da média da população $\bar{Y}_k - \mu$ também é uma variável aleatória normal. Se essa última variável aleatória $(\bar{Y}_k - \mu)$ for dividida pelo desvio-padrão, σ, da população, ainda desconhecido, o resultado é uma variável aleatória normal padrão (média = 0, desvio-padrão = 1). Contudo, não sabemos o desvio-padrão σ da população, e precisamos, em vez disso, dividir os desvios de cada média pela estimativa do erro-padrão da média de cada amostra ($s_{\bar{Y}_k}$). A distribuição-*t* resultante é similar, mas não idêntica a uma normal padrão. Essa distribuição-*t* é leptocúrtica, com caudas mais longas e pesadas que uma normal padrão. [8]

[8] Este resultado foi primeiramente demonstrado pelo estatístico W.S. Gossett, que o publicou usando o pseudônimo "Student". Naquele tempo, Gosset, estava empregado na cervejaria Guinness, que não permitia que seus empregados publicassem segredos comerciais, e, por isso, precisava de um pseudônimo. Esta distribuição normal padrão modificada, chamada por Gossett de distribuição-*t*, também é conhecida como "distribuição-*t* de Student". Conforme o número de amostras aumenta, a forma da distribuição-*t* se aproxima da distribuição normal padrão. A construção da distribuição-*t* requer a especificação do tamanho amostral n e é normalmente escrita como $t_{[n]}$. Para $n = \infty$, $t_{[\infty]} \sim N(0,1)$.

Como a distribuição-*t* para um n pequeno é leptocúrtica (*ver* Figura 3.5), a largura do intervalo de confiança construído a partir dela irá se estreitar conforme o tamanho amostral aumenta. Por exemplo, para $n = 10$, 95% da área sob a curva caem entre ±2,228. Para $n = 100$, 95% da área sob a curva caem entre ±1,990. O intervalo de confiança resultante é 12% mais largo para $n = 10$ que para $n = 100$.

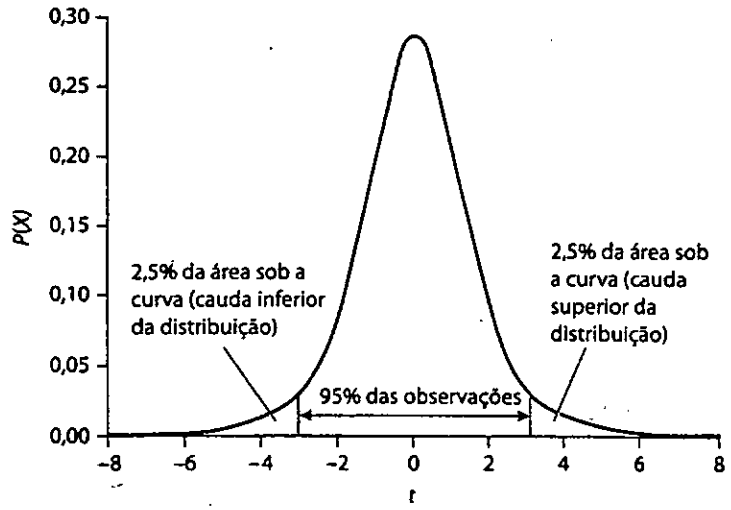


Figura 3.7 Distribuição-*t* ilustrando que 95% das observações, ou massa de probabilidade, caem dentro do percentil de ±1,96 desvio-padrão da média (média = 0). As duas caudas da distribuição contêm cada 2,5% das observações ou massa de probabilidade da distribuição. Elas somam 5% das observações, e a probabilidade $P = 0,05$ de que uma observação caia nessas caudas. Esta distribuição é idêntica à distribuição-*t* ilustrada na Figura 3.5.

RESUMO

Estatísticas descritivas descrevem os valores esperados e a variabilidade de uma amostra aleatória de dados. Medidas de posição incluem a mediana, a moda e diversas médias. Se as amostras são aleatórias ou independentes, a média aritmética é um estimador imparcial do valor esperado de uma distribuição normal. As médias geométrica e harmônica também são utilizadas em circunstâncias especiais. Medidas de dispersão incluem a variância, o desvio-padrão e o erro-padrão, e os quantis. Essas medidas descrevem a variação das observações acerca do valor esperado. Medidas de dispersão também podem ser utilizadas para construir intervalos de confiança ou de credibilidade. As interpretações desses intervalos diferem entre os estatísticos frequentistas e bayesianos.

4

Formulando e Testando Hipóteses

Hipóteses são explicações potenciais que podem representar nossas observações do mundo externo. Em geral, elas representam relações de "causa e efeito" entre um mecanismo ou processo proposto (a causa) e nossas observações (o efeito). Observações são dados – aquilo que vemos ou medimos no mundo real. Nosso objetivo em realizar um estudo científico é entender a(s) causa(s) de fenômenos observáveis. Coletar observações é um meio para este fim: acumulamos diversos tipos de observações e as usamos pra distinguir entre diferentes causas possíveis. Alguns cientistas e estatísticos fazem uma distinção entre observações realizadas durante experimentos manipulativos daquelas realizadas durante estudos **observacionais ou correlacionais.** Contudo, na maioria dos casos, o tratamento estatístico desse tipo de dados é idêntico. A distinção recai sobre as **inferências**[1] feitas com esses estudos. Experimentos manipulativos bem-delineados permitem que fiquemos bastante confiantes sobre nossas inferências; temos menos confiança em dados de experimentos pobremente delineados, ou de estudos em que somos incapazes de manipular diretamente as variáveis.

Se as observações são o "quê" da ciência, as hipóteses são o "como". Enquanto as observações são tiradas do mundo real, as hipóteses não necessariamente o são. Embora nossas observações possam sugerir hipóteses, as hipóteses também podem vir do corpo da literatura científica existente, das predições de modelos teóricos, e de nossa própria intuição e raciocínio. Contudo, nem todas as descrições de relações de "causa e efeito" constituem hipóteses científicas válidas. Uma hipótese científica deve ser testável: em outras palavras, deve haver algum conjunto

[1] Em lógica, uma *inferência* é uma conclusão que é derivada de premissas. Cientistas fazem inferências (tiram conclusões) acerca de causas com base em seus dados. Essas conclusões podem ser sugeridas ou implicadas pelos dados. Mas, lembre-se, é o cientista quem infere o que os dados indicam.

adicional de observações ou resultados experimentais que possamos coletar e que cause a modificação, a rejeição ou o descarte de nossa hipótese de trabalho.[2] Hipóteses metafísicas, incluindo atividades de deuses onipotentes, não se qualificam como científicas porque essas explicações são tiradas da fé, e não há observações que possam fazer um crédulo rejeitá-las.[3]

Além de ser testável, uma boa hipótese científica deve gerar novas predições, as quais podem ser testáveis através da coleta de mais observações. Contudo, o mesmo conjunto de observações pode ser predito por mais de uma hipótese. Embora as hipóteses sejam escolhidas para incorporar nossas observações iniciais, uma boa hipótese científica também deve fornecer um conjunto único de predições que não surjam a partir de outras explicações. Ao manter o foco nessas predições únicas, podemos coletar com mais rapidez os dados críticos que irão discriminar entre as alternativas.

MÉTODOS CIENTÍFICOS

O "método científico" é a técnica usada para decidir entre hipóteses com base em observações e predições. A maioria dos livros-texto apresenta apenas um método científico, mas cientistas praticantes, na verdade, usam vários métodos em seu trabalho.

Dedução e indução

Dedução e indução são dois importantes modos de raciocínio científico, e ambos envolvem fazer inferências a partir de dados ou modelos. A **dedução** vai do caso geral para o específico. O conjunto de afirmações a seguir fornece um exemplo de dedução clássica:

---

[2] Uma *hipótese* científica refere-se a um mecanismo particular ou relação de "causa e efeito"; uma *teoria* científica é muito mais abrangente e mais sintética. Em seus estágios iniciais, nem todos os elementos de uma teoria científica podem ser completamente articulados, de modo que hipóteses explícitas podem não ser possíveis inicialmente. Por exemplo, a teoria da seleção natural de Darwin requeria um mecanismo de herança que conservasse as características de uma geração para a próxima e ao mesmo tempo mantivesse a variação entre indivíduos. Darwin não tinha uma explicação para a herança, e ele discutia essa fraqueza de sua teoria em *The Origin of Species* (1859). Darwin não sabia que tal mecanismo precisamente tinha de fato sido descoberto por Gregor Mendel em seus estudos experimentais (ca. 1856) sobre a herança das ervilhas. Ironicamente, o avô de Darwin, Erasmus Darwin, havia publicado um trabalho sobre herança duas gerações antes, em seu *Zoonomia, or the Laws of Organic Life* (1794-1796). Contudo, Erasmus Darwin usou a planta boca-de-leão como seu organismo de estudo, enquanto Mendel usou ervilhas. A herança da cor das flores é mais simples em plantas de ervilhas que em bocas-de-leão, e Mendel pôde reconhecer a natureza particulada dos genes, enquanto Erasmus Darwin não.

[3] Embora muitas filosofias tenham tentado preencher as lacunas entre ciência e religião, a contradição entre a razão e a fé é uma linha de falha crítica que as separa. Um dos primeiros filósofos do Cristianismo, Tertuliano (~155-222 d.C.) se utilizou dessa contradição e afirmou "*Credo quai absurdum est*" ("Creio porque é absurdo"). Na visão tertuliana, a morte do filho de Deus deve ser acreditada porque é um fato contraditório; e ele ter levantado da tumba é um fato verídico, pois é impossível (Reese, 1980).

1. Todas as formigas de Harvard Forest pertencem ao gênero *Myrmica*.
2. Eu amostrei esta formiga em particular em Harvard Forest.
3. Esta formiga em particular é do gênero *Myrmica*.

As afirmações 1 e 2 geralmente são tratadas como *premissas maior e menor* e a afirmação 3 é a *conclusão*. O conjunto de três afirmações é chamado de **silogismo**, uma importante estrutura lógica desenvolvida por Aristóteles. Note que a sequência do silogismo decorre do caso geral (todas as formigas de Harvard Forest) para o caso específico (a formiga em particular que foi amostrada).

Em contraste, a **indução** vai do caso específico para o geral:[4]

1. Todas estas 25 formigas são do gênero *Myrmica*.
2. Todas estas 25 formigas foram coletadas em Harvard Forest.
3. Todas as formigas de Harvard Forest são do gênero *Myrmica*.

Alguns filósofos definem a dedução como uma inferência certa e a indução como uma inferência provável. Tais definições certamente se ajustam ao exemplo das formigas coletadas em Harvard Forest. No primeiro conjunto de afirmações (dedução), a conclusão precisa ser logicamente verdadeira se as duas premissas são verdadeiras. Mas no segundo caso (indução), embora a conclusão seja provavelmente verdadeira, também pode ser falsa; nossa confiança irá aumentar com o tamanho da amostra, que é sempre o caso em inferência estatística. A estatística, por natureza, é um processo indutivo: sempre tentamos tirar conclusões gerais com base em amostras específicas e limitadas.

Ambas, indução e dedução, são utilizadas em todos os modelos de raciocínio científico, mas elas recebem diferentes ênfases. Mesmo utilizando o método indutivo, provavelmente usaremos a dedução para derivar predições específicas a partir das hipóteses gerais em cada volta do ciclo.

Sir Francis Bacon

[4] O defensor do método indutivo foi Francis Bacon (1561-1626), grande figura jurídica, filosófica e política da Inglaterra de Elizabeth. Ele era um membro proeminente do parlamento, e foi nomeado cavaleiro em 1603. Entre os estudiosos que questionam a autoria dos trabalhos de Shakespeare (os chamados anti-stratfordianos), alguns acreditam que Bacon era o verdadeiro autor das peças de Shakespeare, mas as evidências não são muito convincentes. O trabalho científico mais importante de Bacon foi seu *Novum Organum* (1620), no qual ele incita o uso da indução e do empiricismo como forma de conhecer o mundo. Esta foi uma importante quebra com o passado, em que a exploração da "filosofia natural" envolvia excessiva dependência sobre a dedução e em publicações de autoridades (particularmente os trabalhos de Aristóteles). O método indutivo de Bacon pavimentou o caminho para as importantes descobertas científicas de Galileo e de Newton na Idade da Razão. Perto do fim de sua vida, a fortuna política de Bacon teve uma reviravolta para pior: em 1621, ele foi removido do cargo por aceitar subornos. A devoção de Bacon ao empirismo algumas vezes o levou ao radicalismo. Na tentativa de testar a hipótese de que o congelamento desacelera a putrefação da carne, Bacon se aventurou no frio durante o inverno de 1626 para envolver uma galinha em neve. Ele ficou bastante prejudicado por esse ato e morreu poucos anos depois, aos 65 anos.

Qualquer indagação científica começa com uma observação que tentamos explicar. O método indutivo toma essa observação e desenvolve uma única hipótese para explicá-la. Bacon enfatizava a importância de usar os dados para sugerir a hipótese, ao invés de invocar a sabedoria convencional, as autoridades reconhecidas ou as teorias filosóficas abstratas. Uma vez que a hipótese é formulada, ela gera – através de dedução – predições adicionais. Essas predições são, então, testadas, coletando-se informações adicionais. Se as novas observações se adequarem às predições, a hipótese é apoiada. Caso contrário, a hipótese precisa ser modificada para levar em conta as observações originais e as novas observações. Esse ciclo de hipótese–predição–observação é repetidamente reiniciado. Após, a hipótese modificada deve se aproximar da verdade[5] (Figura 4.1).

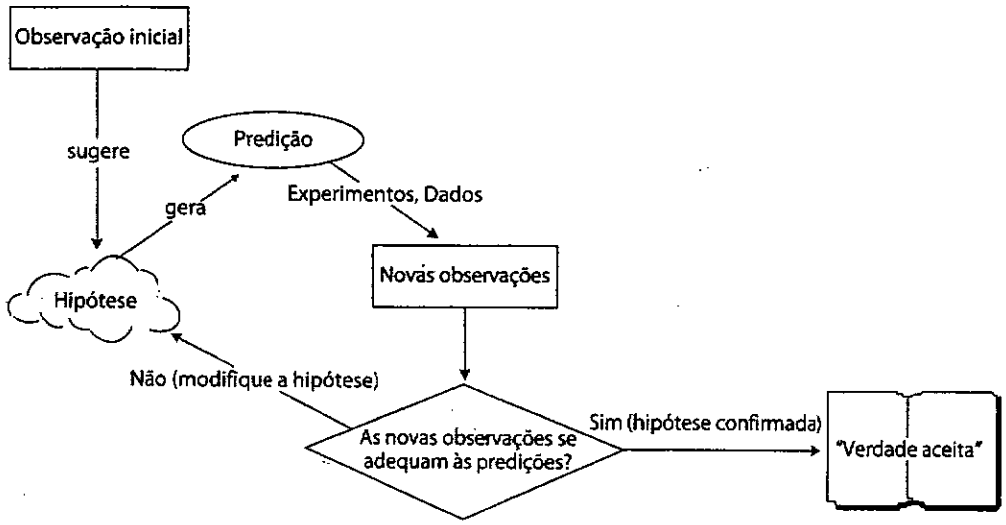


Figura 4.1 O método indutivo. O ciclo de hipótese, predição e observação é repetidamente reiniciado. A confirmação de uma hipótese representa o ponto final teórico do processo. Compare o método indutivo ao hipotético-dedutivo (Figura 4.4), em que múltiplas hipóteses de trabalho são propostas e a ênfase é dada ao falseamento, ao invés da verificação.

[5] Ecólogos e cientistas da área ambiental recorrem à indução quando usam programas estatísticos para ajustar funções não lineares (curvas) aos dados (*ver* Capítulo 9). Os programas requerem que especifique não somente a equação a ser ajustada, mas também um conjunto de valores iniciais para os parâmetros desconhecidos. Esses valores iniciais necessitam estar "próximos" aos reais, pois os algoritmos são estimadores *locais* (i. e., buscam solução para o mínimo ou máximo local da função). Portanto, se as estimativas iniciais estão "distantes" dos valores reais da função, as rotinas de ajuste da curva podem falhar em convergir para uma solução ou irá convergir a uma solução sem sentido algum. Traçar a curva ajustada junto com os dados é uma boa medida de segurança para confirmar que a curva derivada pelos parâmetros estimados realmente se ajusta aos dados originais.

Duas vantagens do método indutivo são (1) enfatizar a íntima ligação entre dados e teoria; e (2) explicitamente construir e modificar hipóteses com base no conhecimento prévio. O método indutivo é *confirmatório* na medida em que buscamos dados que suportem a hipótese, para depois modificar a hipótese para confirmar os dados acumulados.[6]

Existem várias desvantagens no método indutivo. Talvez a mais séria é de que esse método considera somente uma hipótese inicial; outras hipóteses são consideradas apenas depois, em resposta a observações adicionais. Se "partirmos com o pé errado" e começarmos explorando hipóteses equivocadas, pode-se demorar muito para chegar à resposta correta pela indução. Em alguns casos podemos nunca chegar lá. Além disso, o método indutivo pode encorajar cientistas a terem hipóteses "de estimação" e talvez ficarem com elas muito tempo antes de descartá-las ou modificá-las radicalmente (Loehle, 1987). E por fim, o método indutivo – pelo menos na visão de Bacon – deriva teorias exclusivamente a partir de observações empíricas. Contudo, muitas ideias teóricas importantes vieram de modelagem teórica, raciocínio abstrato, e da boa e velha intuição. Hipóteses importantes em todas as ciências frequentemente surgiram em antecipação à coleta dos dados críticos necessários para testá-las.[7]

A indução nos dias modernos: inferência bayesiana

A **hipótese nula** é o ponto inicial de uma investigação científica, pois tenta explicar os padrões nos dados da forma mais simples possível, o que em geral significa atribuir, inicialmente, a variação nos dados à aleatoriedade ou ao erro de medidas. Se essa simples hipótese nula pode ser rejeitada, podemos nos mover

Robert H. MacArthur

[6] O ecólogo de comunidades Robert H. MacArthur (1930-1972) certa vez escreveu que o grupo de pesquisadores interessados em fazer da ecologia uma ciência "organiza dados ecológicos como exemplos para testar as teorias propostas e passa a maior parte do tempo remendando teorias para explicar o maior número de dados quanto possível" (MacArthur, 1962). Essa citação caracteriza muito do trabalho teórico inicial da ecologia de comunidades. Depois, a ecologia teórica se desenvolveu como disciplina por seus próprios méritos e algumas linhas de pesquisa interessantes floresceram sem qualquer referência a dados ou ao mundo real. Os ecólogos discordam acerca de quando um grande corpo de trabalho puramente teórico tem sido bom ou mal para a nossa ciência (Pielou, 1981; Caswell, 1988).

[7] Por exemplo, em 1931, o físico austríaco Wolfgang Pauli (1900-1958) criou a hipótese da existência do neutrino, uma partícula eletricamente neutra com massa desprezível, para explicar inconsistências aparentes na conservação de energia durante o decaimento radioativo. A confirmação empírica da existência do neutrino não aconteceu até 1956.

para uma hipótese mais complexa.[8] Como o método indutivo começa com uma observação que sugere uma hipótese, como gerar uma hipótese nula adequada? A inferência bayesiana representa uma versão moderna e atualizada do método

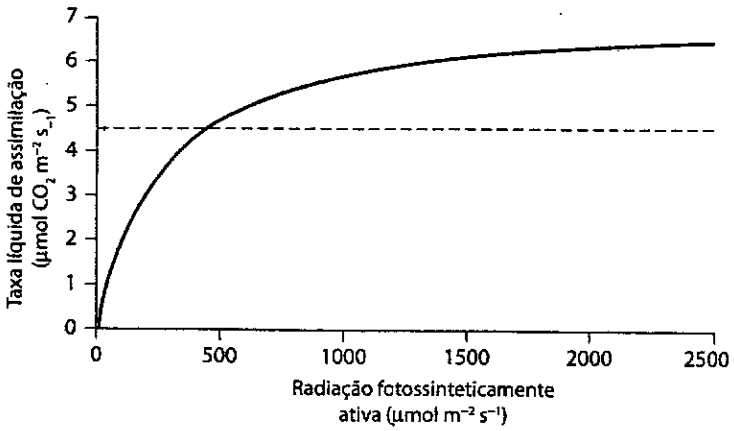


Figura 4.2 Duas hipóteses nulas para a relação entre intensidade luminosa (medida como radiação fotossinteticamente ativa) e a taxa fotossintética (medida como taxa de assimilação líquida) em plantas. A hipótese nula mais simples é de que não existe nenhuma associação entre as duas variáveis (linha pontilhada). A hipótese nula é o ponto de partida para uma abordagem hipotético-dedutiva que assume o conhecimento prévio acerca da relação entre as variáveis e é a base para um modelo de regressão linear (Capítulo 9). Em contraste, a curva verde representa a abordagem bayesiana de trazer o conhecimento prévio para criar uma hipótese nula informativa. Nesse caso, o "conhecimento prévio" vem da fisiologia vegetal e da fotossíntese. Esperamos que a taxa de assimilação cresça rapidamente no começo, conforme a intensidade luminosa aumenta, e então atinja uma assíntota, ou um nível de saturação. Tal relação pode ser descrita pela equação de Michaelis-Menten [$Y = kX/(D + X)$], que inclui os parâmetros para uma taxa de assimilação (k) assintótica e uma constante de meia-saturação (D) que controla a inclinação da curva. Métodos bayesianos podem incorporar este tipo de informação prévia na análise.

Sir William de Ockham

[8] A preferência por hipóteses simples sobre as complexas tem uma longa história na ciência. O Princípio da Parcimônia, de William de Ockham (1290-1349), declara que "[Entidades] não devem ser multiplicadas além da necessidade". Ockham acreditava que as hipóteses muito complexas eram vãs e insultavam Deus. O Princípio da Parcimônia também é conhecido como "Navalha de Ockham", ou seja: a navalha apara a complexidade desnecessária. Ockham viveu uma vida interessante: foi educado em Oxford e membro da Ordem Franciscana. Ele foi acusado de heresia em razão de alguns de seus escritos em sua tese de Mestrado. A acusação foi mais tarde abandonada, mas quando o Papa João XXII desfiou a doutrina Franciscana à pobreza apostólica, Ockham foi excomungado e fugiu para a Bavária. Morreu em 1349, provavelmente vítima da epidemia da peste bubônica.

indutivo. Os fundamentos da inferência bayesiana podem ser ilustrados com um exemplo simples.

A resposta fotossintética das folhas ao aumento da intensidade de luz é um problema bem-estudado. Imagine um experimento em que cultivamos plântulas de mangue-vermelho, cada uma sob uma intensidade diferente de luz (expressa como densidade do fluxo de fótons fotossintéticos, ou DFFF, em fótons µmol por m^2 de tecido de folha exposta à luz por segundo) e medimos a resposta fotossintética de cada planta (expressa como µmoles de CO_2 fixados por m^2 de tecido de folha exposto à luz por segundo). Então representamos graficamente os dados com a intensidade da luz no eixo-*x* (a *variável preditora*) e a taxa fotossintética no eixo-*y* (a *variável resposta*). Cada ponto representa uma folha diferente para a qual registramos esses dois números.

Na ausência de qualquer informação acerca da relação entre a luz e a taxa fotossintética, a hipótese nula mais simples é de que não há relação entre essas duas variáveis (Figura 4.2). Se ajustarmos uma linha a essa hipótese nula, a inclinação da linha seria 0. Se coletarmos dados e encontrarmos alguma outra relação entre a disponibilidade de luz e a taxa fotossintética, podemos, então, usar esses dados para modificar nossa hipótese, seguindo o método indutivo.

Porém, é realmente necessário formular a hipótese nula mesmo que não tenhamos informação nenhuma? Usando apenas um pouco de conhecimento acerca da fisiologia vegetal, podemos formular uma hipótese inicial mais realística. Especificamente, esperamos que exista uma taxa fotossintética máxima que a planta possa atingir. Além desse ponto, o aumento na intensidade luminosa não irá produzir fotossíntese adicional, pois, algum outro fator, como a água ou os nutrientes, se torna limitante. Mesmo que esses fatores sejam fornecidos e a planta seja cultivada sob condições ótimas, a taxa fotossintética ainda estará fora do nível, porque existem limitações inerentes às taxas dos processos bioquímicos e à transferência de elétrons que ocorre durante a fotossíntese. (De fato, se continuarmos aumentando a intensidade luminosa, o excesso de energia luminosa pode danificar o tecido das plantas e reduzir a fotossíntese. Em nosso exemplo, porém, limitamos a amplitude superior da intensidade luminosa àquelas que a planta pode tolerar).

Desse modo, nossa hipótese nula informada é de que a relação entre a taxa fotossintética e a intensidade luminosa deve ser não linear, com uma assíntota as grandes intensidades luminosas (*ver* Figura 4.2). Dados reais poderiam, então, ser utilizados para testar o grau de suporte para essa hipótese mais realística (Figura 4.3). Para determinar qual hipótese nula usar, também precisamos questionar qual, precisamente, é o ponto do estudo? A hipótese nula simples (equação linear) será apropriada se queremos apenas estabelecer que uma relação não aleatória existe entre a intensidade luminosa e a taxa fotossintética. A hipótese nula informada (equação de Michaelis-Menten) será apropriada se

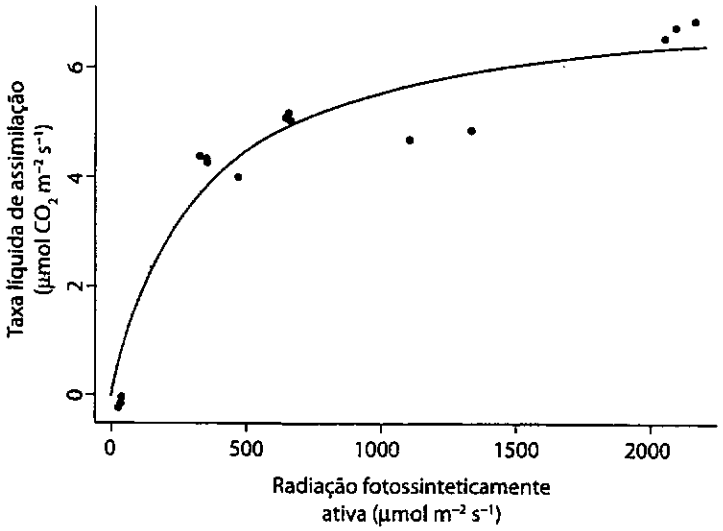


Figura 4.3 Relação entre a intensidade luminosa e a taxa fotossintética. Os dados são medidas da taxa de assimilação líquida e da radiação fotossinteticamente ativa para $n = 15$ folhas jovens da planta do mangue-vermelho *Rhizophora mangle* em Belize (Farnsworth e Ellison, 1996b). A equação de Michaelis-Menten com a forma $Y = kX/(D + X)$ foi ajustada aos dados. As estimativas dos parâmetros ±1 desvio-padrão são $k = 7{,}3 \pm 0{,}59$ e $D = 313 \pm 86{,}6$.

desejarmos comparar curvas de saturação entre espécies ou testar modelos teóricos que fazem predições quantitativas para a assíntota ou para a constante de meia-saturação.

As Figuras 4.2 e 4.3 ilustram como um indutivista dos dias atuais, ou estatístico bayesiano, constroi uma hipótese. A abordagem bayesiana é de usar informação ou conhecimento prévio para gerar e testar hipóteses. Neste exemplo, o conhecimento prévio é derivado da fisiologia vegetal e da forma esperada da curva de saturação luminosa. Contudo, o conhecimento prévio também pode ser fundamentado na extensa base de literatura publicada sobre curvas de saturação luminosa (Björkman, 1981; Lambers et al., 1998). Se tivéssemos estimativas empíricas de parâmetros de outros estudos, poderíamos quantificar nossas estimativas prévias do valor limiar e da assíntota para a saturação luminosa. Essas estimativas poderiam, então, ser usadas para especificar a hipótese inicial sobre o ajuste do valor da assíntota aos nossos dados experimentais.

O uso de conhecimento prévio dessa forma é diferente da de Bacon, da indução, que se baseava inteiramente na própria experiência do indivíduo. Em um universo baconiano, se você nunca tiver estudado plantas, não pode ter nenhuma evidência direta sobre a relação entre intensidade luminosa e taxa fotossintética, e começaria com uma hipótese nula, como a linha horizontal da Figura 4.2. Eis o ponto inicial do método hipotético-dedutivo, que será apresentado na próxima sessão.

A interpretação estritamente baconiana da indução é a base fundamental das críticas à abordagem bayesiana: o conhecimento prévio utilizado para desenvolver o modelo inicial é arbitrário e subjetivo, e pode ser enviesada pelas noções preconcebidas do pesquisador. Desse modo, o método hipotético-dedutivo é visto por alguns como sendo mais "objetivo" e, portanto, mais científico. Os bayesianos rebatem o argumento, afirmando que a hipótese nula estatística e as técnicas de ajuste de curvas utilizadas por hipotético-dedutivistas são igualmente subjetivas; esses métodos apenas parecem ser mais objetivos por serem familiares e acriticamente aceitos. Para uma discussão adicional desses temas filosóficos, *ver* Ellison (1996, 2004) e Dennis (1996).

O método hipotético-dedutivo

O **método hipotético-dedutivo** (Figura 4.4) desenvolveu-se a partir dos trabalhos de Isaac Newton e outros cientistas do século XVII e foi defendido pelo filósofo da ciência Karl Popper.[9] Como o método indutivo, o hipotético-dedutivo começa com uma observação inicial que estamos tentando explicar. Contudo, em vez de criar uma única hipótese nula e prosseguir, esse método nos faz propor múltiplas hipóteses de trabalho. Todas consideram a observação inicial, mas cada uma faz predições adicionais que podem ser testadas com experimentos e observações adicionais. O objetivo desses testes não é confirmar, mas falsear as hipóteses. O falseamento elimina algumas explicações e a lista é reduzida a um número menor de hipóteses competidoras. O ciclo de predições e novas observações é repetido. O método hipotético-dedutivo, porém, nunca confirma uma hipótese; a explicação cientificamente aceita é a hipótese que resiste com sucesso a repetidas tentativas de falseá-la.

As duas vantagens do método hipotético-dedutivo são: (1) ele força a consideração de múltiplas hipóteses de trabalho desde o início; e (2) ele destaca as diferenças preditivas-chave entre elas. Em contraste ao método indutivo, as hipóteses não necessitam ser construídas a partir de dados, mas podem ser desenvolvidas independentemente ou em paralelo à coleta de dados. A ênfase no falseamento tende a produzir hipóteses simples e testáveis, de forma que as

Karl Popper

[9] O austríaco e filósofo da ciência Karl Popper (1902-1994) foi o mais articulado defensor do método hipotético-dedutivo e da falseabilidade como pedra-chave da ciência. Em *The logic of Scientific Discovery* (1935), Popper demonstrou que a falseabilidade é um critério mais confiável do que a verificabilidade. Em *The open Science and Its Enemies* (1945), ele defendeu a democracia e criticou as implicações totalitárias da indução e as teorias políticas de Platão e Karl Marx.

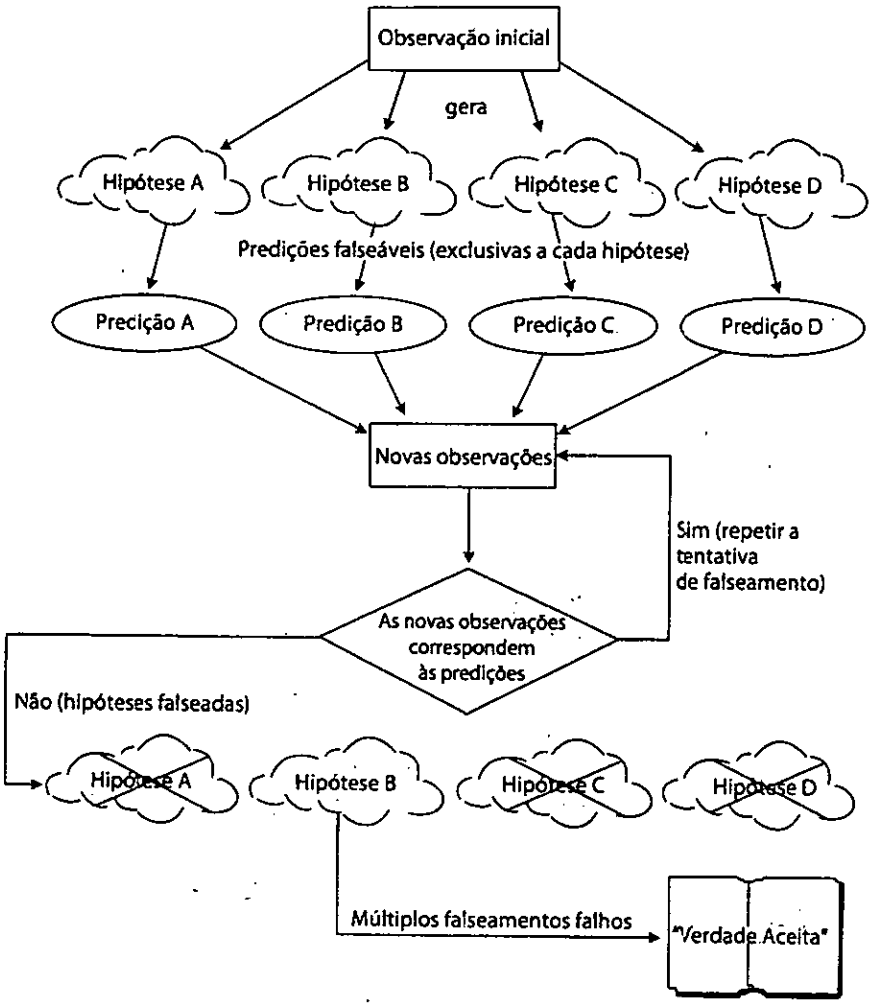


Figura 4.4 O método hipotético-dedutivo. Hipóteses múltiplas de trabalho são propostas e suas predições são testadas com o objetivo de falsear as incorretas. A explicação correta é aquela que se mantém depois de repetidos testes que falham em falseá-la.

explicações mais cautelosas são consideradas primeiro, e os mecanismos mais complicados por último.[10]

[10] A árvore da lógica é uma variante bem conhecida do método hipotético-dedutivo com a qual você deve estar familiarizado dos cursos de química. A árvore da lógica diz respeito à decisão dicotômica, em que diferentes ramos são seguidos dependendo do resultado de experimentos em cada bifurcação da árvore. A dica do ramo terminal representa as diferentes hipóteses que estão sendo testadas. A árvore da lógica também pode ser encontrada nas familiares-chaves taxonômicas dicotômicas usadas para identificar espécies de plantas ou animais desconhecidas: "Se o animal tem 3 pares de pernas

(*Continua*)

As desvantagens do método hipotético-dedutivo são de que múltiplas hipóteses de trabalho nem sempre estão disponíveis, particularmente nos estágios iniciais da investigação. Mesmo se múltiplas hipóteses estiverem disponíveis, o método não funciona, a menos que a hipótese "correta" esteja entre as alternativas. Em contraste, o método indutivo pode começar com uma hipótese incorreta, mas pode chegar à explicação correta através de repetidas modificações da hipótese original, como consequência da coleta de dados. Outra distinção útil é que o método indutivo ganha força ao comparar muitos conjuntos de dados a uma única hipótese, enquanto o método hipotético-dedutivo é melhor para comparar um único conjunto de dados à múltiplas hipóteses. Por fim, ambos os métodos indutivo e hipotético-dedutivo enfatizam uma única hipótese correta, dificultando a avaliação de casos em que múltiplos fatores estão atuando. Esse é um problema que afeta menos a abordagem indutiva, pois múltiplas explicações podem ser incorporadas a hipóteses mais complexas.

Nenhum método científico é o correto, e alguns filósofos da ciência negam que nenhum cenário realmente descreve como a ciência opera.[11] Contudo, os métodos hipotético-dedutivo e indutivo caracterizam muito da ciência no mundo real (em oposto ao mundo abstrato da filosofia da ciência). A razão para investir tempo nesses modelos é para entender suas relações com os testes estatísticos de hipóteses.

---

[10] (*Continuação*)

locomotoras, vá para o passo *x*; se ele tem 4 ou mais pares, vá para o passo *y*." A árvore da lógica nem sempre é prática para hipóteses ecológicas complexas; pode haver excessivos ramos, e eles podem não ser todos dicotômicos. Contudo, ela sempre é um excelente exercício para criar e expor suas idéias e seus experimentos em uma estrutura compreensível. Platt (1964) defendeu esse método e apontou o seu sucesso espetacular na biologia molecular; a descoberta da estrutura helicoidal do DNA é um exemplo clássico do método hipotético-dedutivo (Watson e Crick, 1953).

Thomas Kuhn

[11] Nenhuma discussão de Popper e do método hipotético-dedutivo poderia ser completa sem mencionar a nêmese filosófica de Popper, Thomas Kuhn (1922-1996). Em *The Structure of Scientific Revolutions* (1962), Kuhn colocou em questão toda a estrutura dos testes de hipóteses, e justificou que eles não representam a forma como a ciência foi concebida. Kuhn acreditava que ela havia sido concebida dentro do contexto de **paradigmas** principais, ou estruturas de pesquisas, e que o domínio desses paradigmas foi implicitamente adotado em cada geração de cientistas. As atividades dos cientistas de "solucionar o quebra-cabeça" constituem a "ciência ordinária", na qual anomalias empíricas são conciliadas aos paradigmas existentes. Entretanto, nenhum paradigma pode abranger todas as observações, e conforme as anomalias se acumulam, o paradigma se torna pouco efetivo. Eventualmente, ele entra em colapso e há uma revolução científica em que é substituído por um paradigma inteiramente novo. Assumindo uma instância relativamente intermediária entre Popper e Kuhn, o filósofo Imre Lakatos (1922-1974) pensava que os programas de pesquisas científicas (PPC) consistiam em um núcleo de princípios centrais que geravam um cinturão de hipóteses subjacentes para fazer mais predições específicas. As predições de uma hipótese podem ser testadas por um método científico, mas o núcleo não é diretamente acessível (Lakatos, 1978). Um paralelo entre Kuhn, Popper, Lakatos e outros filósofos da ciência pode ser lido em Lakatos & Musgrave (1970). *Ver também* Horn (1986) para uma discussão adicional dessas ideias.

TESTANDO HIPÓTESES ESTATÍSTICAS

Hipótese estatística *versus* retorno à hipótese científica

O uso da estatística para testar hipóteses é somente uma pequena faceta do método científico, mas ele consome uma quantidade desproporcional do nosso tempo, bem como espaço em revistas. Usamos estatística para descrever padrões em nossos dados, e depois testes estatísticos para decidir se as predições de uma hipótese são suportadas ou não. O estabelecimento de hipóteses, a articulação de suas predições, o delineamento e a execução de experimentos válidos, além da coleta, a organização e a descrição dos dados ocorrem antes do uso dos testes estatísticos. Enfatizamos que aceitar ou rejeitar uma hipótese estatística é bem diferente de fazer o mesmo com uma hipótese científica. A hipótese nula estatística em geral é do tipo "não há padrão", como não há diferença entre grupos, ou não há relação entre duas variáveis contínuas. Em contraste, a hipótese alternativa é de que o padrão existe. Em outras palavras, existem diferenças distintas em valores medidos entre grupos, ou uma relação evidente existe entre duas variáveis contínuas. Você pode se questionar como esses padrões estão relacionados às hipóteses científicas que você está testando.

Por exemplo, imagine que você está avaliando a hipótese científica de que a arrebentação de ondas na costa rochosa cria um espaço vazio em razão da remoção de espécies de invertebrados competitivamente dominantes. O espaço aberto pode ser colonizado por espécies competitivamente subordinadas que em outro caso seriam excluídas. Essa hipótese prediz que a diversidade de espécies de invertebrados marinhos irá mudar em função do nível de distúrbio (Sousa, 1979). Você coleta dados sobre o número de espécies em superfícies rochosas afetadas e não afetadas por distúrbios. Usando um teste de hipóteses adequado, você observa que não há diferença na riqueza de espécies entre os dois grupos. Neste caso, você *falhou em rejeitar* a hipótese nula estatística, e o padrão dos dados *falhou em dar suporte* a uma das predições da hipótese de distúrbios. Note, contudo, que a ausência de evidência não é evidência de ausência; falhar em rejeitar a hipótese nula não é equivalente a aceitar a hipótese nula (apesar de frequentemente ser tratada como se fosse).

Aqui está um segundo exemplo no qual o padrão estatístico é o mesmo, mas a conclusão científica é diferente. A distribuição ideal livre é uma hipótese que prediz que os organismos se movem entre hábitats e ajustam sua densidade de forma que tenha a mesma aptidão média em diferentes hábitats (Fretwell e Lucas, 1970). Uma predição testável dessa hipótese é que a aptidão dos organismos em hábitats diferentes é similar, mesmo pensando que a densidade da população possa diferir. Suponha que você mede a taxa de crescimento populacional de pássaros (um componente importante da aptidão de aves) em hábitats de florestas e em descampados como um teste para esta predição (Gill et al., 2001). Como no primeiro exemplo, você falha em rejeitar a hipótese nula estatística, e por isso não

há evidência de que as taxas de crescimento diferem entre os hábitats. Mas, nesse caso, *falhar em rejeitar* a hipótese nula estatística de fato *dá suporte* à predição da distribuição ideal livre.

Naturalmente, existem muitas observações adicionais e testes que podemos precisar para poder avaliar a hipótese do distúrbio ou a distribuição ideal livre. O ponto é que as hipóteses estatística e científica são entidades distintas. Em qualquer estudo, é necessário determinar quando o ato de dar suporte ou refutar a hipótese nula estatística fornece evidência negativa ou positiva para a hipótese científica. Tal determinação também influencia profundamente a forma como você estipula seu estudo experimental ou protocolo amostral para estudos observacionais. A distinção entre hipótese nula estatística e a hipótese científica é tão importante que retornaremos a ela mais adiante neste capítulo.

Significância estatística e valores de *P*

É praticamente universal reportar os resultados de um teste estatístico para afirmar a importância dos padrões que observamos nos dados que coletamos. Uma afirmação típica é: "Os grupos controle e o tratamento diferiram significativamente um do outro ($P = 0,01$)". O que, precisamente, um "$P = 0,01$" significa, e como se relaciona aos conceitos de probabilidades apresentados nos Capítulos 1 e 2?

UM EXEMPLO HIPOTÉTICO: COMPARANDO MÉDIAS Um problema de avaliação comum nas ciências ambientais é determinar se as atividades humanas resultam em um aumento do estresse de animais. Em vertebrados, o estresse pode ser medido através dos níveis dos hormônios glicocorticóides (GC) no sangue ou nas fezes. Por exemplo, lobos que não são expostos a carros de neve têm 872,0 ng GC/g, enquanto os expostos a carros de neve têm 1.468,0 ng GC/g (Creel et al., 2002). Agora, como decidir se essa diferença é grande suficiente para ser atribuída à presença dos carros de neve?[12]

Aqui é onde você pode conduzir um teste estatístico convencional. Tais testes podem ser muito simples (como o familiar teste-*t*), ou mais complexos (com

[12] Muitas pessoas tentam responder esta questão simplesmente comparando as médias. Contudo, não podemos avaliar a diferença entre médias a menos que tenhamos uma ideia do quanto os indivíduos dentro de um mesmo grupo diferem. Por exemplo, se muitos indivíduos do grupo sem carros de neve têm níveis de GC tão baixos quanto 200 ng/g e outros com níveis de GC tão altos quanto 1544 ng/g (a média, lembre-se, foi de 872), então uma diferença de 600 ng/g entre os dois grupos pode não significar muito. Por outro lado, se a maioria dos indivíduos do grupo sem carros de neve tem níveis de GC entre 850 e 950 ng/g, então uma diferença de 600 ng/g seria substancial. Como discutimos no Capítulo 3, precisamos saber não somente a diferença entre as médias, mas a variância acerca delas – a quantidade que um indivíduo típico difere da média de seu grupo. Sem saber algo sobre a variância não podemos dizer nada sobre se as diferenças entre as médias dos dois grupos têm algum significado.

os de termos de interação em uma análise de variâncias). Porém, todos os testes estatísticos desse tipo produzem como resultado uma **estatística do teste**, que é somente o resultado numérico do teste, e um **valor de probabilidade** (ou **valor de *P***), que é associado à estatística do teste.

A HIPÓTESE NULA ESTATÍSTICA Antes de podermos definir a probabilidade de um teste estatístico, precisamos definir a hipótese nula estatística ou H_0. Vimos, que os cientistas preferem explicações cautelosas ou simples ao invés das mais complexas. Qual é a explanação mais simples para explicar a diferença nas médias dos dois grupos? Em nosso exemplo, a mais simples é de que as diferenças representam variação aleatória entre os grupos e não refletem qualquer efeito sistemático dos carros de neve. Em outras palavras, se dividirmos os lobos em dois grupos, mas não expormos os indivíduos de nenhum deles aos carros de neve, ainda assim podemos encontrar médias que diferem uma da outra. Lembre-se que é extremamente improvável que as médias de duas amostras de números sejam iguais, mesmo se elas forem amostradas a partir de uma população maior, usando um processo idêntico.

Os níveis de glicocorticóides irão diferir de um indivíduo para outro por muitas razões que não podem ser estudadas ou controladas neste experimento, e toda esta variação – incluindo aquela devido a erros de medida – é o que rotulamos de variação aleatória. Queremos saber se há qualquer evidência de que a diferença na média observada nos níveis de GC dos dois grupos é maior que esperaríamos dada a variação aleatória entre os indivíduos. Assim, uma hipótese nula estatística típica seria que "diferenças entre os grupos não são maiores que a que esperaríamos devido à variação aleatória." Chamamos isso de **hipótese nula estatística**, porque a hipótese é de que algum mecanismo ou força específicos – alguma força *que não* a variação aleatória – *não* está operando.

A HIPÓTESE ALTERNATIVA Uma vez que declaramos a hipótese nula estatística, definimos então uma ou mais *alternativas* à hipótese nula. Em nosso exemplo, a **hipótese alternativa** natural é de que a diferença observada na média dos níveis de GC dos dois grupos é muito grande para ser explicada apenas pela variação aleatória entre indivíduos. Note que a hipótese alternativa *não* é de que a exposição aos carros de neve é responsável pelo aumento de GC! Em vez disso, a hipótese alternativa é focada simplesmente no *padrão* presente nos dados. O pesquisador pode *inferir* mecanismos a partir dos padrões, mas essa inferência é um passo à parte. O teste estatístico meramente revela se o padrão é provável ou não, dado que a hipótese nula seja verdadeira. Nossa habilidade em apontar mecanismos causais a esses padrões estatísticos depende da qualidade do nosso delineamento experimental e das nossas medidas.

Por exemplo, suponha que o grupo de lobos expostos aos carros de neve também tenham sido caçados e perseguidos por humanos e seus cachorros de caça durante o dia anterior, enquanto o grupo não exposto inclui lobos de uma área remota não habitada por humanos. As análises estatísticas provavelmente revelariam diferenças significativas nos níveis de GC entre os dois grupos independentemente da exposição aos carros de neve. Contudo, seria arriscado concluir que as diferenças entre as médias dos dois grupos foi causada pelos carros de neve, mesmo pensando que podemos rejeitar a hipótese nula estatística de que o padrão é explicado pela variação aleatória entre indivíduos. Neste caso, o efeito do tratamento é **confundido** com outras diferenças entre os grupos-controle e tratamento (exposição aos cachorros de caça), que são potencialmente relacionadas aos níveis de estresse. Conforme discutiremos nos Capítulos 6 e 7, um objetivo importante para um bom delineamento amostral é evitar confusões.

Se nosso experimento fosse delineado e executado corretamente, seria seguro inferir que a diferença entre as médias é causada pela presença de carros de neve. Porém, mesmo assim, não podemos apontar precisamente o mecanismo fisiológico se tudo que fizemos foi medir os níveis de GC de indivíduos expostos e não expostos. Precisaríamos de informações muito mais detalhadas sobre a fisiologia hormonal, química, sanguínea e similares, se quisermos chegar aos mecanismos subjacentes.[13] A estatística nos ajuda a estabelecer padrões convincentes e a partir deles podemos fazer inferências ou tomar conclusões acerca de relações de causa e efeito.

Na maioria dos testes, a hipótese alternativa não é explicitamente declarada porque em geral há mais de uma que pode explicar os padrões dos dados. Em vez disso, consideramos o conjunto de alternativas como sendo "não H_0". Em um diagrama de Venn, todos os resultados de dados podem então ser classificados ou em H_0 ou não H_0.

O VALOR DE *P* Em muitas análises estatísticas queremos saber quando a hipótese nula de variação aleatória entre indivíduos pode ser rejeitada. O valor de *P* é um guia para tomar essa decisão. Um valor de *P* estatístico mede a probabilidade de que a diferença observada ou mais extrema poderia ser encontrada *se a hipótese nula fosse verdadeira.* Utilizando a notação de probabilidades condicionais apresentada no Capítulo 1, valor de $P = P(\text{dados} \mid H_0)$.

[13] Mesmo se os mecanismos fisiológicos forem elucidados, ainda haveria questões acerca dos mecanismos finais no nível molecular ou genético. Sempre que propomos um mecanismo haverá processos em níveis menores que não são completamente descritos por nossas explicações e precisam ser tratados como uma "caixa preta." Contudo, nem todos os processos em níveis maiores podem ser explicados com sucesso pelo reducionismo a mecanismos de níveis menores.

Suponha que o valor de *P* seja relativamente pequeno (próximo de 0,0). Então é improvável (a probabilidade é pequena) que a diferença observada tenha sido obtida se a hipótese nula fosse verdadeira. No exemplo dos lobos e carros de neve, um valor de *P* baixo pode significar que é improvável que a diferença de 600 ng/g no nível de GC tenha sido observada entre os grupos expostos e não expostos caso houvesse apenas variação aleatória entre os indivíduos e nenhum efeito consistente dos carros de neve (i. e., se a hipótese nula fosse verdadeira). Portanto, com um valor de *P* pequeno, o resultado seria improvável dada à hipótese nula, então a rejeitamos. Como tínhamos apenas uma hipótese alternativa em nosso estudo, nossa conclusão é de que carros de neve (ou algo associado a eles) podem ser responsáveis pela diferença entre os grupos de tratamentos.[14]

[14] Aceitar uma hipótese alternativa com base neste mecanismo de testar a hipótese nula é um exemplo da falácia de "afirmar o consequente" (Baker, 1989). Formalmente, o valor de $P = P(\text{dados} \mid H_0)$. Caso a hipótese nula seja verdadeira, poderia resultar em (ou nos termos da lógica, *implicar*) um conjunto particular de observações (neste caso, os dados). Podemos escrever isso formalmente como $H_0 \Rightarrow$ dados nulos, onde a seta é lida como "implica." Se suas observações são diferentes das esperadas sob H_0, então um valor de *P* baixo sugere que $H_0 \not\Rightarrow$ seus dados, e a seta cortada é lida como "não implica". Como você estipulou apenas uma hipótese alternativa, H_a, então você está estipulando também que $H_a = \neg H_0$ onde o símbolo $\neg$ significa "não", e as únicas possibilidades para os dados são os possíveis sob H_0 ("dados nulos") e aqueles que não são possíveis sob H_0 ("$\neg$ dados nulos" = "seus dados"). Assim, você está declarando a seguinte lógica de progressão:

1. Dado: $H_0 \Rightarrow$ dados nulos
2. Observe: $\neg$ dados nulos
3. Conclua: $\neg$ dados nulos $\Rightarrow \neg H_0$
4. Assim: $\neg H_0$ ($=H_a$) $\Rightarrow \neg$ dados nulos

Mas, realmente, tudo que você pode concluir é o ponto 3: $\neg$ dados nulos $\Rightarrow \neg H_0$ (o conhecido *contrapositivo* de 1). Em 3, a hipótese alternativa (H_a) é o "consequente," e você não pode declarar sua verdade simplesmente observando seu "predicado" ($\neg$ dados nulos em 3); muitas outras causas possíveis poderiam produzir esses resultados ($\neg$ dados nulos). Você pode afirmar o consequente (declarar H_a como verdadeira) se e somente se houver *somente uma* alternativa possível à sua hipótese nula. No caso mais simples, onde H_0 declara "nenhum efeito" e H_a, "algum efeito," proceder de 3 para 4 faz sentido. Mas biologicamente, em geral é mais interessante saber qual é o verdadeiro efeito (oposto a simplesmente mostrar que há "algum efeito").

Considere o exemplo anterior, das formigas, neste capítulo. Permita que H_0 = todas as 25 formigas em Harvard Forest são *Myrmica*, e H_a = 10 formigas na floresta não são *Myrmica*. Caso colete um espécime de *Camponotus* na floresta, você pode concluir que os dados implicam que a hipótese nula é falsa (observação de uma *Camponotus* na floresta $\Rightarrow \neg H_0$). Mas você não pode tirar nenhuma conclusão sobre a hipótese alternativa. Você pode apoiar uma hipótese alternativa menos rigorosa, H_a = nem todas as formigas na floresta são *Myrmica*, mas afirmar essa hipótese alternativa não lhe dirá nada sobre a verdadeira distribuição das formigas na floresta, ou a identidade das espécies e dos gêneros que estão presentes.

Eis um ponto vital. Muitos cientistas parecem acreditar que, quando reportam um valor de *P*, estão fornecendo a probabilidade de observar a hipótese nula conforme os dados [$P(H_0 \mid \text{dados})$]. Ou a probabilidade de que a hipótese alternativa seja falsa, de acordo com os dados [$1 - P(H_a \mid \text{dados})$]. Mas, de fato, eles estão reportando algo completamente diferente – a probabilidade de observar os dados de acordo com a hipótese nula: $P(\text{dados} \mid H_0)$. Infelizmente, como vimos no Capítulo 1, $P(\text{dados} \mid H_0) \neq P(H_0 \mid \text{dados}) \neq 1 - P(H_a \mid \text{dados})$; nas palavras de um filósofo anônimo imortal de Maine, *você não pode chegar lá a partir daqui*. No entanto, é possível calcular diretamente $P(H_0 \mid \text{dados})$ ou $P(H_a \mid \text{dados})$ usando o Teorema de Bayes (Capítulo 1) e os métodos esboçados no Capítulo 5.

Por outro lado, suponha que o valor de P calculado seja relativamente grande (próximo a 1,0). Então é provável que a diferença observada possa ter ocorrido, dado que a hipótese nula é verdadeira. Neste exemplo, um valor de *P* grande pode significar que uma diferença de 600-ng/g nos níveis de GC provavelmente poderia ser observada entre os grupos expostos e não expostos mesmo quando os carros de neve não tiverem nenhum efeito e que haja apenas variação aleatória entre os indivíduos. Ou seja, com um valor de *P* grande, os resultados observados poderiam ser prováveis sob a hipótese nula, então não temos evidência suficiente para rejeitá-la. Nossa conclusão é que as diferenças nos níveis de GC entre os dois grupos pode ser atribuída à variância aleatória entre indivíduos.

Tenha em mente que quando calculamos a um valor de *P* estatístico estamos olhando os dados através das lentes da hipótese nula. Se os padrões em nossos dados são prováveis sob a hipótese nula (valor de *P* alto), não temos nenhuma razão para rejeitá-la em favor de explicações mais complexas. Por outro lado, se os padrões são improváveis sob a hipótese nula (valor de *P* baixo), é mais cauteloso rejeitá-la e concluir que algo mais, além de variação aleatória entre indivíduos, contribui para os resultados.

O QUE DETERMINA O VALOR DE *P*? O valor de *P* calculado depende de três coisas: do número de observações nas amostras (n); da diferença entre as médias das amostras ($\bar{Y}_i - \bar{Y}_j$); e do nível de variação entre indivíduos (s^2). Quanto mais observações em uma amostra, menor o valor de *P*, pois, quanto mais dados temos, mais provável estimarmos a verdadeira média da população e podermos detectar uma diferença real entre elas, caso ela exista (*ver* Lei dos Grandes Números, Capítulo 3). O valor de *P* também será menor quanto mais diferentes os dois grupos forem em relação à variável que estamos medindo. Desse modo, uma diferença de 10 ng/g na média do nível de GC entre os grupos-controle e tratamento irá gerar um valor de *P* mais baixo que uma diferença de 2 ng/g, mantendo as outras coisas iguais. Por fim, o valor de *P* será menor caso a variância entre indivíduos dentro de cada tratamento for pequena. Quanto menos variação houver de um indivíduo para o próximo, mais fácil será detectar diferenças entre os grupos. No caso extremo, se os níveis de GC de todos indivíduos dentro do grupo de lobos expostos aos carros de neve forem idênticos, então qualquer diferença nas médias dos dois grupos, não importa quão pequena seja, pode gerar um valor de *P* baixo.

QUANDO UM VALOR DE *P* É SUFICIENTEMENTE BAIXO? Em nosso exemplo, obtemos um valor de $P = 0{,}01$ para a probabilidade de obter a diferença observada nos níveis de GC entre lobos expostos e não expostos a carros de neve. Desse modo, se a hipótese nula for verdadeira e houver apenas variação aleatória entre os indivíduos nos dados, a chance de encontrar uma diferença de 600 ng/g na GC entre os grupos expostos e não expostos é de somente 1 em 100. Em outras palavras, se a hipótese nula for verdadeira e conduzirmos esse experimento 100 vezes, usando diferentes indivíduos em cada, poderíamos esperar que em somente *um* dos experimentos víssemos uma diferença tão grande ou maior que a que de fato ob-

servamos. Portanto, parece improvável que a hipótese nula seja verdadeira, então a rejeitamos. Se nosso experimento foi delineado de maneira adequada, podemos concluir, com segurança, que os carros de neve causam aumento nos níveis de GC, apesar de não podermos especificar o quê, nos carros de neve, causa esta resposta. Por outro lado, se o valor da probabilidade estatística calculada fosse $P = 0,88$, poderíamos esperar um resultado similar ao que encontramos em 88 de 100 experimentos, devido à variação aleatória entre os indivíduos; assim, o resultado observado poderia não ser de todo incomum sob a hipótese nula e não haveria razão para rejeitá-la.

Mas qual é o ponto de corte preciso que devemos empregar para rejeitar ou não uma hipótese nula? Eis uma decisão subjetiva, já que não existe nenhum valor natural crítico abaixo do qual devemos sempre rejeitar a hipótese nula e acima do qual nunca devemos rejeitá-la. Entretanto, após muitas décadas de prática, tradição, e de um esforço vigilante dos editores e revisores de periódicos, o valor operacional crítico para tomar essa decisão é de 0,05. Ou seja, se a probabilidade estatística $P \leq 0,05$, a convenção diz para rejeitar a hipótese nula; e se a probabilidade estatística $P > 0,05$, ela não é rejeitada. Quando cientistas reportam que um resultado em particular é "significativo", quer dizer que rejeitaram a hipótese nula com um valor de $P \leq 0,05$.[15]

Um pouco de reflexão deve convencê-lo de que um valor crítico de 0,05 é relativamente baixo. Se usar essa regra em seu cotidiano, você nunca levaria um guarda-chuva consigo a menos que a previsão de chuva seja, no mínimo, de 95%. Você ficaria ensopado muito mais frequentemente que seus vizinhos e amigos. Por outro lado, se eles o virem carregando um guarda-chuva, podem ficar bastante confiantes de que vai mesmo chover.

Em outras palavras, assumir um valor crítico = 0,05 como padrão para rejeitar a hipótese nula é muito conservador. Isso requer que a evidência seja muito forte para rejeitar a hipótese nula estatística. Alguns pesquisadores estão descontentes com a utilização de um valor crítico arbitrário e por assumir um valor tão baixo quanto 0,05. Afinal, a maioria de nós pegaria um guarda-chuva com uma previsão de 90%; então, por que não deveríamos ser um pouco menos rígidos em nosso padrão para rejeitar a hipótese nula? Talvez devêssemos assumir um valor crítico = 0,10, ou talvez usar valores críticos diferentes para variadas formas de dados e questões.

[15] Quando cientistas discutem resultados "significativos" em seus trabalhos, estão na verdade falando sobre quão confiantes estão de que a hipótese nula tenha sido rejeitada corretamente. Mas o público iguala "significativo" a "importante". Contudo, essa distinção não causa um fim à confusão, e é uma das razões pelas quais os cientistas gastam bastante tempo comunicando claramente suas ideias em revistas populares.

Uma defesa do ponto de corte de 0,05 é a observação de que os padrões científicos precisam ser altos, de forma que os pesquisadores possam se apoiar com confiança no trabalho alheio. Se a hipótese nula for rejeitada a um padrão mais liberal, haverá maior risco de rejeitar uma hipótese nula verdadeira (erro Tipo I, descrito em mais detalhes a seguir). Se estivermos tentando construir hipóteses e teorias científicas com base nos dados e resultados de outros, tais erros diminuem o progresso científico. Ao usar um valor crítico baixo, podemos ficar confiantes de que os padrões nos dados são, de certo modo, fortes. Contudo, mesmo um valor crítico baixo não é uma proteção contra estudos ou experimentos maldelineados. Em tais casos, a hipótese nula pode ser rejeitada, mas os padrões nos dados refletem falhas na amostragem ou nas manipulações, e não relacionadas às diferenças biológicas subjacentes que estamos tentando entender.

Talvez o argumento mais forte a favor da necessidade de um valor crítico baixo é o de que nós, humanos, somos psicologicamente predispostos a reconhecer e ver padrões em nossos dados, mesmo quando inexistentes. Nosso sistema sensorial vertebrado é adaptado para organizar dados e observações em padrões "úteis", gerando um viés embutido no sentido de rejeitar a hipótese nula e de ver padrões onde, na verdade, há aleatoriedade (Sale, 1984).[16] Um valor crítico baixo é uma medida de segurança contra tais atividades e que também ajuda a agir como uma barreira sobre a taxa de publicações científicas, pois é muito menos provável que os resultados não significativos sejam reportados ou publicados.[17] Enfatizamos, contudo, que nenhuma lei *impõe* um valor crítico $\leq 0,05$ para que os resultados sejam declarados significantes. Em muitos casos, pode ser mais útil reportar o exato valor do P e deixar que os leitores decidam por si só o quanto os resultados são importantes. No entanto, na realidade prática, revisores e editores habitualmente não permitirão que você discuta os mecanismos que não são apoiados por um resultado com $P \leq 0,05$.

[16] Uma ilustração fascinante disto é pedir a um amigo para desenhar um conjunto de 25 pontos localizados de forma aleatória em um pedaço de papel. Se você comparar a distribuição desses pontos a um conjunto de pontos realmente aleatórios gerados por um computador, você verá com frequência que os pontos desenhados são distintamente não aleatórios. As pessoas tendem a espaçar os pontos de forma muito equidistante pelo papel, enquanto um padrão verdadeiramente aleatório gera "agregados" e "buracos" aparentes. Em razão dessa tendência a ver padrões em tudo, devemos usar um valor crítico baixo para garantir que não estamos enganando a nós mesmos.

[17] A conhecida tendência das revistas em rejeitarem artigos com resultados não significantes (Murtaugh, 2002a) – e, consequentemente, os autores que não se esforçaram para tentar publicá-los –, não é algo positivo. No método hipotético-dedutivo, a ciência avança através da eliminação de hipóteses alternativas, e isso com frequência pode ser feito quando falhamos em rejeitar a hipótese nula. Entretanto, essa abordagem requer que os autores especifiquem e testem as predições que são feitas pelas hipóteses alternativas competidoras. Testes estatísticos com base em H_0 *versus* sem H_0 frequentemente não permitem essa forma de especificidade.

HIPÓTESE ESTATÍSTICA *VERSUS* REDUCIONISMO DE HIPÓTESES CIENTIFÍCAS A maior dificuldade em usar um valor de *P* resulta da falha em distinguir a hipótese nula estatística das científicas. Lembre-se de que uma *hipótese científica* representa um mecanismo formal para explicar os padrões nos dados. Neste caso, nossa hipótese científica é a de que carros de neve causam estresse em lobos, o que propusemos testar medindo os níveis de GC. Altos níveis de GC podem surgir de complexas mudanças na fisiologia e levar a mudanças na produção de GC quando um animal está estressado. Em contraste, a *hipótese nula estatística* é uma declaração acerca dos padrões nos dados e da verossimilhança com que esses padrões podem surgir por acaso ou por processos aleatórios que não são relacionados de maneira explícita aos fatores que estamos estudando.

Usamos os métodos de probabilidade quando decidimos se rejeitamos ou não a hipótese nula estatística; pense nesse processo como uma maneira de padronizar os dados. A seguir, concluímos acerca da validade de nossa hipótese científica baseando-nos nos padrões estatísticos dos dados. A força desta inferência depende muito dos detalhes dos delineamentos experimental e amostral. Em um experimento bem-delineado e replicado, que inclui controles apropriados e que os indivíduos são atribuídos aleatoriamente a tratamentos bem-definidos, podemos ficar bastante confiantes acerca de nossas inferências e de nossa habilidade para avaliar a hipótese científica sob consideração. Contudo, em um estudo amostral no qual não podemos manipular nenhuma variável, mas simplesmente medir diferenças entre grupos, torna-se difícil fazer inferências sólidas acerca da hipótese científica subjacente, mesmo se tivermos rejeitado a hipótese nula estatística.[18]

Acreditamos que a questão mais geral não é o valor crítico em particular que é escolhido, mas se sempre devemos usar uma estrutura de testes de hipóteses. Com certeza, para muitas questões, os testes de hipóteses estatísticos são uma poderosa forma de estabelecer quais padrões existem ou não nos dados. Porém, em muitos estudos, a verdadeira questão pode não ser o teste de hipóteses, mas a **estimativa de parâmetros**. Por exemplo, no caso do estudo de estresse, pode ser mais importante determinar a amplitude dos níveis de GC esperada para os lobos expostos aos carros de neve, em vez de apenas estabelecer que os carros de neve aumentam de forma significativa os níveis de GC.

[18] Em contraste ao exemplo dos carros de neve e dos lobos, suponha que tenhamos medido os níveis de GC de 10 lobos velhos escolhidos de maneira aleatória e de 10 lobos jovens escolhidos da mesma forma. Poderíamos ficar tão confiantes acerca de nossas inferências, como no caso do experimento dos carros de neve? Por quê? Quais são as diferenças, se houver, entre experimentos em que manipulamos indivíduos em diferentes grupos (lobos expostos *versus* não expostos) e amostragens em que medimos a variação entre grupos, mas não manipulamos diretamente ou alteramos as suas condições (lobos velhos *versus* jovens)?

Além disso, também deveríamos estabelecer o nível de confiança ou certeza em nossa estimativa de parâmetros.

Erros em testes de hipóteses

Apesar de a estatística envolver muitos cálculos precisos, é importante não perder de vista o fato de que a estatística é uma disciplina mergulhada em incertezas. Tentamos usar dados limitados e incompletos para fazer inferências acerca de mecanismos subjacentes que podemos entender apenas em parte. Na realidade, a hipótese nula estatística é verdadeira ou falsa; se tivermos informações perfeitas e completas, podemos saber quando ela é verdadeira ou não, e não precisaríamos de estatística para afirmar isso. Ao contrário, temos apenas nossos dados e métodos de inferência estatística para decidir quando rejeitar ou não a hipótese nula estatística. Isso leva a uma interessante tabela 2×2 de resultados possíveis sempre que testamos uma hipótese nula estatística (Tabela 4.1).

Idealmente, gostaríamos de encerrar nas células superior esquerda ou na inferior direita da Tabela 4.1. Em outras palavras, quando há apenas variação aleatória em nossos dados, esperaríamos não rejeitar a hipótese nula estatística (célula superior esquerda) e, quando houver algo mais, rejeitá-la (célula inferior direita). Contudo, podemos nos encontrar em uma das outras duas células, que correspondem a dois tipos de erros que podem ser cometidos em uma decisão estatística.

ERRO TIPO I Se rejeitarmos erroneamente uma hipótese nula que era verdadeira (célula superior direita da Tabela 4.1), teremos feito uma afirmação falsa de que algum fator acima e além da variação aleatória está causando o padrão

TABELA 4.1 O mundo quádruplo dos testes de hipóteses

	Manter H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta	Erro Tipo I (α)
H_0 falsa	Erro Tipo II (β)	Decisão correta

A hipótese nula subjacente é verdadeira ou falsa, mas no mundo real precisamos usar amostragens e dados limitados para tomar a decisão de aceitar ou não a hipótese nula. Sempre que uma decisão estatística é tomada, um entre quatro resultados possíveis irá ocorrer. Uma decisão correta ocorre quando mantemos uma hipótese nula que é verdadeira (canto superior esquerdo), ou rejeitamos uma hipótese nula que é falsa (canto inferior direito). As outras duas possibilidades representam erros no processo de decisão. Se rejeitarmos uma hipótese nula verdadeira, cometemos o erro Tipo I (canto superior direito). Testes paramétricos padrão buscam controlar o α, a probabilidade do erro Tipo I. Se mantivermos uma hipótese nula falsa, cometemos o erro Tipo II (canto inferior esquerdo). Então, a probabilidade de erro Tipo II é β.

em nossos dados. Este é o erro Tipo I, e por convenção, a probabilidade de cometê-lo é denotada por α. Quando você calcula um valor de *P* estatístico, está, na verdade, estimando o α. Desse modo, uma definição mais precisa do valor de *P* é de que ele representa a chance de que iremos cometer um erro Tipo I, de erroneamente rejeitar uma hipótese nula verdadeira.[19] Essa definição dá mais apoio para aceitar a significância estatística somente quando o valor de *P* é muito pequeno. Quanto menor o valor de *P*, mais confiantes podemos ficar de que não cometeremos um erro Tipo I se rejeitarmos H_0. No exemplo do glicocorticoide, o risco de cometer um erro Tipo I ao rejeitar a hipótese nula é de 1%. Como vimos anteriormente, publicações científicas têm um padrão de um risco máximo de 5% de erro Tipo I de rejeitar a hipótese nula. Na avaliação de impactos ambientais, o erro Tipo I pode ser um "falso-positivo" no qual, por exemplo, o efeito de um poluente sobre a saúde humana é reportado, mas de fato não existe.

ERRO TIPO II E PODER ESTATÍSTICO A célula inferior da esquerda, na Tabela 4.1, representa o erro Tipo II. Neste caso, o investigador falha incorretamente em rejeitar uma hipótese nula que é falsa. Em outras palavras, existem diferenças sistemáticas entre os grupos sendo comparados, mas o investigador falha em rejeitar a hipótese nula e conclui – incorretamente – que há somente a variação aleatória entre as observações. Por convenção, a probabilidade de cometer um erro Tipo II é denotada por β. Na avaliação ambiental, o erro Tipo II pode ser um "falso-negativo" no qual, por exemplo, há o efeito de um poluente sobre a saúde humana, mas ele não é detectado.[20]

[19] Seguimos tratamentos estatísticos padrão que equiparam o valor de *P* calculado com a estimativa da taxa α de erro Tipo I. Contudo, o valor evidente do *P* de Fisher pode não ser estritamente equivalente ao α de Neyman e Pearson. Os estatísticos discordam se essa distinção é uma importante questão filosófica ou simplesmente uma diferença semântica. Hubbard e Bayarri (2003) argumentam que a incompatibilidade é importante, e o artigo deles é seguido por discussões, comentários e respostas de outros estatísticos. Fique por dentro!

[20] A relação entre os erros Tipo I e Tipo II permeiam as discussões do princípio de precaução em tomadas de decisões ambientais. Ao longo da história, por exemplo, agências de controle assumem que novos produtos químicos são benignos até que provem o contrário. São necessárias evidências muito fortes para rejeitar a hipótese nula de que não há efeito na saúde e no bem-estar. Os fabricantes de químicos e outros poluentes em potencial estão empenhados em minimizar a probabilidade de cometer um erro Tipo I. Em contraste, grupos ambientalistas que representam o público geral estão interessados em minimizar a probabilidade de que os fabricantes cometam o erro Tipo II. Tais grupos assumem que um químico é nocivo até que prove ser benigno, e estão dispostos a aceitar uma probabilidade maior em cometer um erro Tipo I, se isso significa que podem ficar mais confiantes de o fabricante não aceitar erroneamente a hipótese nula. Seguindo tal raciocínio, ao avaliar o controle de qualidade da produção industrial, os erros Tipo I e Tipo II são com frequência conhecidos por erros do produtor e do consumidor, respectivamente (Sokal e Rolf, 1995).

Um conceito relacionado à probabilidade de cometer o erro Tipo II é o **poder** do teste estatístico. O poder é calculado como $1 - \beta$ e equipara a probabilidade de corretamente rejeitar à hipótese nula, quando falsa. Queremos que nossos testes estatísticos tenham poder para que haja um aumento nas chances de detectar padrões significativos em nossos dados, quando presentes.

QUAL É A RELAÇÃO ENTRE ERRO TIPO I E TIPO II? Como ideal, gostaríamos de minimizar tanto o erro Tipo I quanto o Tipo II em nossas inferências estatísticas. Contudo, estratégias designadas para reduzir o erro Tipo I inevitavelmente aumentam o risco de erro Tipo II, e vice-versa. Por exemplo, suponha que decida rejeitar a hipótese nula somente se $P < 0,01$ – um padrão 5 vezes mais rigoroso que o critério convencional de $P < 0,05$. Embora seu risco de cometer o erro Tipo I seja agora muito menor, existe uma chance muito maior de que, quando falhar em rejeitar a hipótese nula, poderá fazê-lo incorretamente (i. e., você estará cometendo o erro Tipo II). Apesar dos erros Tipo I e Tipo II serem inversamente relacionados um ao outro, não existe nenhuma relação matemática entre eles, pois a probabilidade do erro Tipo II depende, em parte, de qual é a hipótese alternativa, de quão grande é o efeito que queremos detectar (Figura 4.5), do tamanho amostral e da sagacidade do nosso delineamento experimental ou protocolo amostral.

POR QUE AS DECISÕES ESTATÍSTICAS SE BASEIAM NO ERRO TIPO I? Em contraste à probabilidade de cometer o erro Tipo I, que determinamos com testes estatísticos padrão, a probabilidade de cometer o erro Tipo II não é reportada ou calculada com frequência, e em muitos artigos científicos a probabilidade de cometer o erro Tipo II, nem é discutida. Por que não? Para começar, muitas vezes não podemos calcular a probabilidade de erro Tipo II, a menos que a hipótese alternativa seja especificada de maneira completa. Em outras palavras, se queremos determinar o risco de erroneamente aceitar a hipótese nula, as alternativas devem ser desenvolvidas alem de "não H_0". Em contraste, calcular a probabilidade de um erro Tipo I não requer essa especificação; em vez disso, é necessário apenas cumprir alguns pressupostos de normalidade e independência (*ver* Capítulos 9 e 10).

Em uma base filosófica, alguns autores argumentaram que o erro Tipo I é um equívoco mais sério na ciência do que o do Tipo II (Shrader-Frechette e McCoy, 1992). O erro Tipo I é de falsidade, no qual incorretamente rejeitamos a hipótese nula e alegamos um mecanismo mais complexo. Outros podem seguir nosso trabalho e tentar construir seus próprios estudos com base naquela alegação falsa. Em contraste, o de Tipo II é um erro de ignorância. Apesar de não termos rejeitado a hipótese nula, alguém com um experimento melhor ou com mais dados poderá ter condição de rejeitá-la no futuro, e a ciência irá progredir daquele ponto. No entanto, é em muitos problemas aplicados, como o monitoramento ambiental

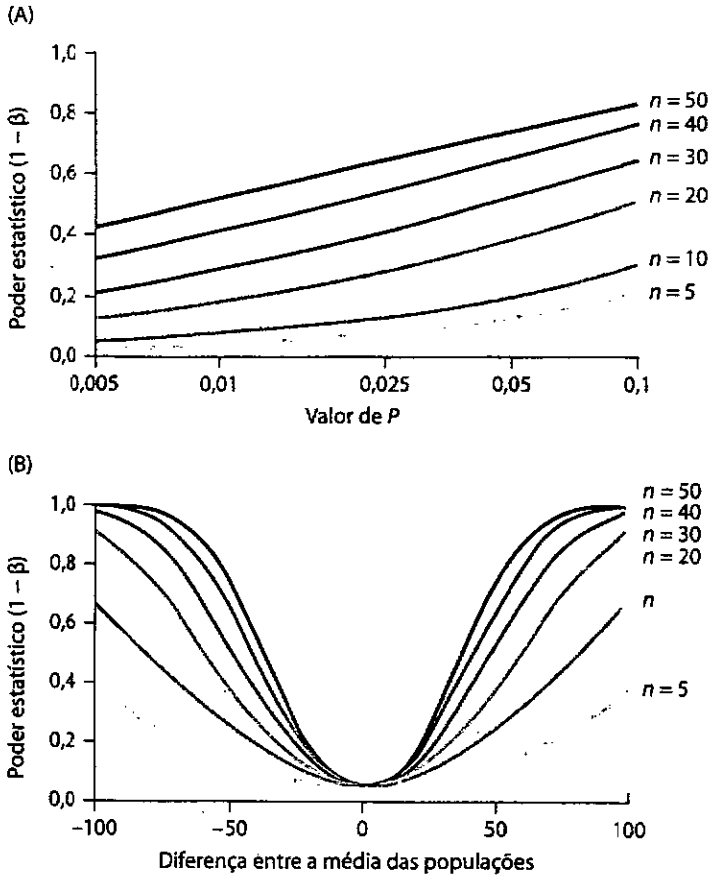


Figura 4.5 Relação entre poder estatístico, valor de *P* e tamanho do efeito observável em função do tamanho amostral. (A) O valor de *P* é a probabilidade de incorretamente rejeitar uma hipótese nula verdadeira, enquanto o poder estatístico é a probabilidade de corretamente rejeitar uma hipótese nula falsa. O resultado geral propõe que, quanto menor o valor de *P* usado para rejeitar a hipótese nula, menor o poder estatístico de corretamente detectar um efeito do tratamento. A um dado valor de *P*, o poder estatístico é maior quando o tamanho amostral é maior. (B) Quanto menor o tamanho do efeito observável do tratamento (i. e., quanto menor a diferença entre o grupo-tratamento e o grupo-controle), maior é o tamanho amostral necessário a um bom poder estatístico para detectar o efeito do tratamento.[21]

[21] Podemos aplicar esses gráficos ao nosso exemplo da comparação dos níveis do hormônio glicocorticoide das populações de lobos expostos a carros de neve (grupo-tratamento) *versus* lobos que não foram expostos aos carros de neve (grupo-controle). Nos dados originais (Creel et al., 2002), o desvio-padrão da população de lobos controle não exposta a carros de neve foi 73,1 e o dos expostos aos carros de neve foi de 114,2. O painel (A) sugere que se houvesse uma diferença de 50 ng/g entre as populações experimentais e que se o tamanho amostral fosse $n = 50$ em cada grupo, os pesquisadores poderiam corretamente aceitar a hipótese alternativa somente em 51% das vezes para um $P = 0{,}01$. O painel (B) mostra que ocorre um aumento súbito no poder conforme as populações se tornam mais diferentes. No estudo real e bem concebido (Creel et al., 2002), o tamanho amostral foi de 193 no grupo não exposto e de 178 no exposto; as diferenças entre as médias das populações foi de 598 ng/g; e o poder estatístico real do teste foi próximo de 1,0 para um $P = 0{,}01$.

e o diagnóstico de doenças, os erros Tipo II podem ter consequências mais graves, pois doenças ou efeitos ambientais adversos poderiam não ser corretamente detectados.

ESTIMATIVA DE PARÂMETROS E PREDIÇÃO

Todos os métodos para teste de hipóteses que descrevemos – o método indutivo (e seu descendente moderno, a inferência bayesiana), o método hipotético-dedutivo e os testes estatísticos de hipóteses – estão centrados na escolha de uma única "resposta" explanatória a partir de um conjunto inicial de múltiplas hipóteses. Na ecologia e nas ciências ambientais, é mais provável que diversos mecanismos podem operar simultaneamente para produzir os padrões observados; um arcabouço de testes de hipóteses que enfatiza explicações únicas pode não ser apropriado. Em vez de testar hipóteses múltiplas, pode ser mais proveitoso estimar a contribuição relativa de cada uma para um padrão em particular. Esta abordagem é esboçada na Figura 4.6, na qual fazemos a partição dos efeitos de cada hipótese sobre os padrões observados, estimando o quanto cada causa contribui para o efeito observado.

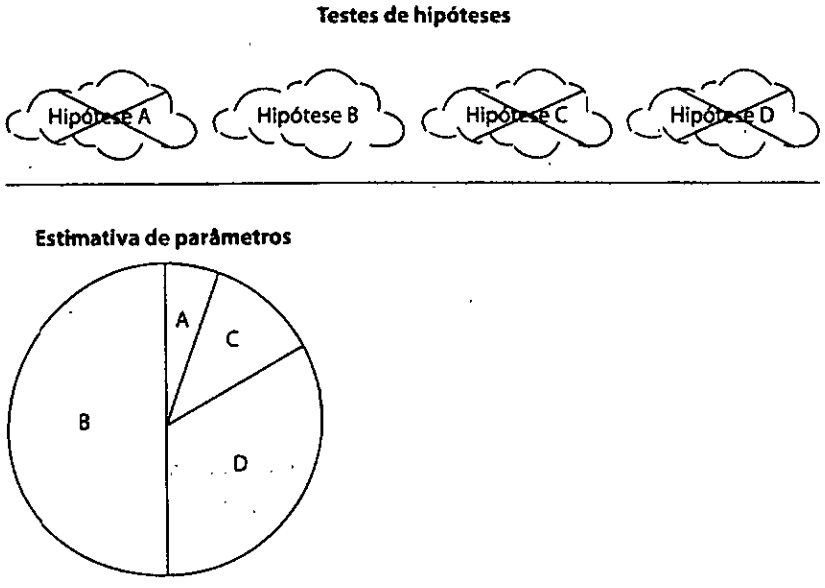


Figura 4.6 Teste de hipóteses *versus* estimativa de parâmetros. A estimativa de parâmetros acomoda com mais facilidade mecanismos múltiplos e pode permitir uma estimativa da importância relativa dos diferentes fatores. A estimativa de parâmetros pode envolver a construção de intervalos de credibilidade (*ver* Capítulo 3) para estimar a força de um efeito. Uma técnica relacionada, na análise de variância, é decompor a variação total dos dados em proporções que são explicadas por diferentes fatores no modelo (*ver* Capítulo 10). Ambos os métodos quantificam a importância relativa de diferentes fatores, enquanto o teste de hipóteses enfatiza uma decisão binária de sim/não sobre se um fator tem um efeito mensurável ou não.

Em tais casos, ao invés de questionar quando uma causa em particular tem algum efeito *versus* nenhum efeito (i. e., ele é significativamente diferente de 0,0?), questionamos qual é a melhor estimativa do **parâmetro** que expressa a magnitude do efeito.[22] Por exemplo, na Figura 4.3, as taxas fotossintéticas medidas para as folhas jovens de *Rhizophora mangle* se ajustam à equação de Michaelis-Menten. Essa equação é um modelo simples que descreve uma variável subindo aos poucos para uma assíntota. A equação de Michaelis-Menten aparece com frequência na biologia, sendo utilizada para descrever desde cinética enzimática (Real, 1977) até taxas de forrageamento de invertebrados (Holding, 1959).

A equação de Michaelis-Menten tem a seguinte forma

$$Y = \frac{kX}{X+D}$$

onde k e D são os dois parâmetros ajustados do modelo, e X e Y são as variáveis independente e dependente. Neste exemplo, k representa a assíntota da curva – que, neste caso, é a taxa máxima de assimilação; a variável independente X é a intensidade luminosa; e a variável dependente Y é a taxa líquida de assimilação. Para os dados da Figura 4.3, a estimativa do parâmetro k é uma taxa máxima de assimilação de 7,1 μmol CO_2 m^{-2} s^{-1}. Isto está de acordo com uma "estimativa visual" de onde a assíntota poderia estar no gráfico.

O segundo parâmetro da equação de Michaelis-Menten é o D, a constante de meia saturação. Esse parâmetro fornece o valor da variável X em que o rendimento da variável Y é a metade da assíntota. Quanto menor o D, mais rápido a curva chega à assíntota. Para os dados da Figura 4.3, a estimativa para o parâmetro D é a radiação fotossinteticamente ativa (RFA) de 250 μmol CO_2 m^{-2} s^{-1}.

Também podemos medir a incerteza na estimativa dos parâmetros através da estimativa do erro-padrão, para construir intervalos de credibilidade (*ver* Capítulo 3). O erro-padrão estimado para $k = 0,49$ e para o $D = 71,3$. Os testes de hipóteses e a estimativa de parâmetros são relacionados, pois, se o intervalo de confiança da incerteza inclui o 0,0, em geral não teremos condições para rejeitar a hipótese nula de nenhum efeito para um dos mecanismos. Para os parâmetros k e D da Figura 4.3, o valor de P para o teste da hipótese nula de que o parâmetro não é diferente de 0,0 é de 0,0001 e 0,004, respectivamente. Desse modo, podemos ficar bastante confiantes em nossa declaração de que esses parâmetros são maiores que 0,0. Mas, para a proposta de avaliação e ajuste de modelos, os valores numéricos dos parâmetros são mais informativos do que somente questionar se eles diferem ou não de 0,0. Nos próximos capítulos daremos outros exemplos de estudos nos quais os parâmetros do modelo são estimados a partir dos dados.

[22] O Capítulo 9 apresenta algumas das estratégias utilizadas para ajustar curvas e estimar parâmetros de dados. *Ver* Hilborn e Mangel (1997) para uma discussão detalhada. Clark et al. (2003) descrevem estratégias bayesianas recentes para ajuste de curvas.

RESUMO

A ciência é feita a partir dos métodos indutivos e hipotético-dedutivos. Em ambos, os dados observados são comparados aos preditos pelas hipóteses. Pelo método indutivo, que inclui as modernas análises bayesianas, uma única hipótese é testada e modificada repetidas vezes; o objetivo é confirmar ou declarar a probabilidade de uma hipótese em particular. Em contraste, o método hipotético-dedutivo requer o estabelecimento simultâneo de múltiplas hipóteses. Ambos são testados contra observações, com objetivo de falsear ou eliminar quase todas hipóteses alternativas, restando uma. Estatísticas são usadas para testar as hipóteses com objetividade, e podem ser usadas tanto nas abordagens indutivas quanto nas hipotético-dedutivas.

As probabilidades são calculadas e reportadas com virtualmente todos os testes estatísticos. Os valores de probabilidade associados a esses testes podem nos permitir inferir causas para o fenômeno em estudo. Testes de hipóteses estatísticas usando o método hipotético-dedutivo fornecem estimativas relacionadas à chance de obter um resultado igual ou mais extremo que o observado, dado que a hipótese nula seja verdadeira. O valor de *P* também é a probabilidade de rejeitar incorretamente a hipótese nula (ou de cometer o erro estatístico do Tipo I). Por convenção e tradição, 0,05 é o ponto de corte nas ciências para dizer que um resultado é estatisticamente significante. O cálculo do valor de *P* depende do número de observações, da diferença entre as médias dos grupos que estão sendo comparados e da quantidade de variação entre indivíduos dentro de cada grupo. O erro estatístico do Tipo II ocorre quando uma hipótese nula falsa é erroneamente aceita. Este tipo de erro pode ser tão sério quanto o do Tipo I, mas a probabilidade de erro Tipo II raras vezes é reportada em publicações científicas. Os testes de hipóteses estatísticas usando métodos indutivos ou bayesianos produzem estimativas da probabilidade de uma hipótese ou hipóteses de interesse, conforme os dados observados. Como são métodos confirmatórios, não fornecem estimativas dos erros Tipo I e Tipo II. Em vez disso, os resultados são expressos como possibilidades (*odds*) ou como a verossimilhança de que uma hipótese em particular seja correta.

Independente do método usado, toda ciência progride articulando hipóteses testáveis, coletando dados que podem ser usados para testar as predições das hipóteses e relacionando os resultados às relações de causa e efeito subjacentes.

5

Três Estruturas Para Análises Estatísticas

Neste capítulo apresentaremos as três maiores estruturas para análises estatísticas: **análises de Monte Carlo, análises paramétricas e análises bayesianas**. Resumidamente, a primeira faz pressupostos mínimos acerca da distribuição subjacente dos dados. Ela usa randomizações dos dados observados como base para as inferências. As análises paramétricas assumem que os dados foram amostrados de uma distribuição subjacente com forma conhecida, como aquelas descritas no Capítulo 2, e estimam os parâmetros da distribuição a partir dos dados. Estimam, ainda, probabilidades da frequência observada dos eventos e usam essas probabilidades como base para as inferências. Como consequência, é um tipo de inferência frequentista. As análises bayesianas também assumem que os dados foram amostrados de uma distribuição subjacente de forma conhecida. E estimam parâmetros não apenas dos dados, mas também a partir do conhecimento prévio, e atribuem probabilidades a esses parâmetros. Essas probabilidades formam as bases da inferência bayesiana. A maioria dos textos convencionais ensina estatística paramétrica aos estudantes, mas as outras duas são também importantes, e as análises de Monte Carlo, na verdade, são mais fáceis de aprender inicialmente. Para apresentar esses métodos iremos usar cada um deles para analisar o mesmo problema amostral.

5.1 O PROBLEMA AMOSTRAL

Imagine que você está tentando comparar a densidade de formigueiros do tipo forrageadoras em dois hábitats – campo e floresta. Neste problema amostral, não nos preocuparemos com a hipótese científica em teste (talvez ainda nem tenhamos uma neste ponto); simplesmente seguiremos pelo processo de obter e analisar os dados para determinar se existem diferenças consistentes na densidade de formigueiros nos dois hábitats.

Você visita uma floresta e um campo adjacente e estima a densidade média de formigueiros em cada um, usando amostragem replicada. Em cada hábitat você escolhe uma localização ao acaso, coloca uma parcela quadrada com área de 1 m^2 e, com cuidado, conta todos os formigueiros que há dentro da parcela. Você repete esse procedimento muitas vezes em cada hábitat. A questão da escolha de locais aleatórios é muito importante para qualquer tipo de análise estatística. Sem métodos mais complicados, como na amostragem estratificada, a randomização é a única medida de segurança para garantir uma amostra representativa da população (*ver* Capítulo 6).

A escala espacial da amostragem determina o escopo da inferência. Estritamente falando, esse delineamento amostral permitirá que você discuta diferenças entre a densidade de formigas em floresta e em campo *neste sítio em particular*. Um melhor delineamento seria visitar várias florestas e campos diferentes e amostrar uma parcela dentro de cada um. Assim, as conclusões poderiam ser mais facilmente generalizadas.

A Tabela 5.1 apresenta os dados em uma **planilha.** Eles estão dispostos com os nomes das linhas e colunas. Cada coluna da tabela indica uma variável diferente que foi medida ou registrada para cada observação. Neste caso, sua intenção original foi a de amostrar parcelas em 6 campos e 6 florestas, mas as parcelas em campos consumiam mais tempo que o esperado e você conseguiu coletar apenas 4 amostras em campos. Dessa forma, a tabela de dados tem 10 linhas (além da primeira, que apresenta os nomes), pois você coletou 10 amostras diferentes (6 de florestas e 4 de campos). Há 3 colunas. A primeira contém um número ID exclusivo atribuído a cada réplica. As outras colunas contêm as duas informações

TABELA 5.1 Planilha de amostras usada para ilustrar as análises de Monte Carlo, paramétricas e bayesianas

Número ID	Hábitat	Número de formigueiros por parcela
1	Floresta	9
2	Floresta	6
3	Floresta	4
4	Floresta	6
5	Floresta	7
6	Floresta	10
7	Campo	12
8	Campo	9
9	Campo	12
10	Campo	10

Cada linha é uma observação independente. A primeira coluna identifica a réplica com um número ID único; a segunda indica o hábitat amostrado; e a terceira, o número de formigueiros registrados em cada réplica.

registradas de cada réplica: o hábitat no qual foi amostrada (campo ou floresta); e o número de formigueiros registrados em cada parcela.[1]

Seguindo os procedimentos do Capítulo 3, calcule a média e o desvio-padrão para cada amostra (Tabela 5.2). Represente graficamente os dados usando um gráfico de barras convencional das médias e dos desvios-padrão (Figura 5.1), ou o mais informativo *box plot* descrito no Capítulo 3 (Figura 5.2).

Apesar de os números coletados nos campos e em florestas mostrarem alguma sobreposição, eles parecem formar dois grupos distintos: as amostras de floresta com média de 7,0 formigueiros por parcela, e as amostras em campos com média de 10,75 formigueiros por parcela. Por outro lado, nosso tamanho amostral é muito pequeno ($n = 6$ amostras em florestas e $n = 4$ em campos). Talvez essas diferenças tenham surgido apenas por acaso ou por amostragem aleatória. Precisamos fazer um teste estatístico antes de decidir se essas diferenças são significativas ou não.

TABELA 5.2 Estatística descritiva para os dados amostrais da Tabela 5.1

Hábitat	*N*	Média	Desvio-padrão
Floresta	6	7,00	2,19
Campo	4	10,75	1,50

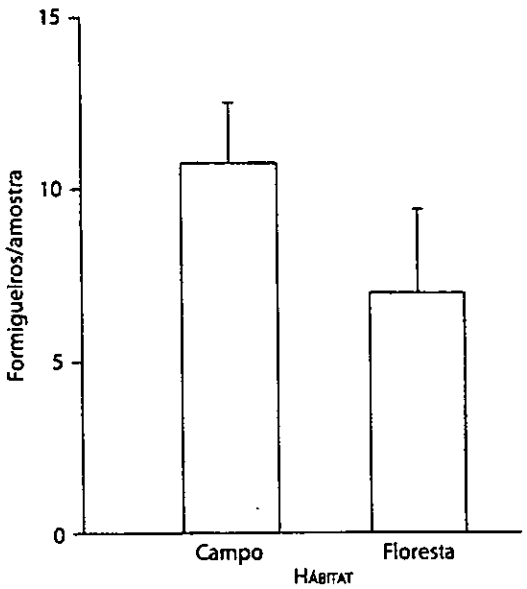


Figura 5.1 Gráfico de barras padrão para os dados amostrais da Tabela 5.1. A altura da barra é a média da amostra e a linha vertical indica um desvio-padrão acima da média. As análises de Monte Carlo, paramétrica e bayesiana podem ser usadas para avaliar a diferença nas médias entre os grupos.

[1] Muitos textos de estatística poderiam mostrar esses dados como duas colunas de números, uma para floresta e outra para campo. Contudo, a forma que mostramos é, em geral, a reconhecida pelos *softwares* estatísticos para análises de dados.

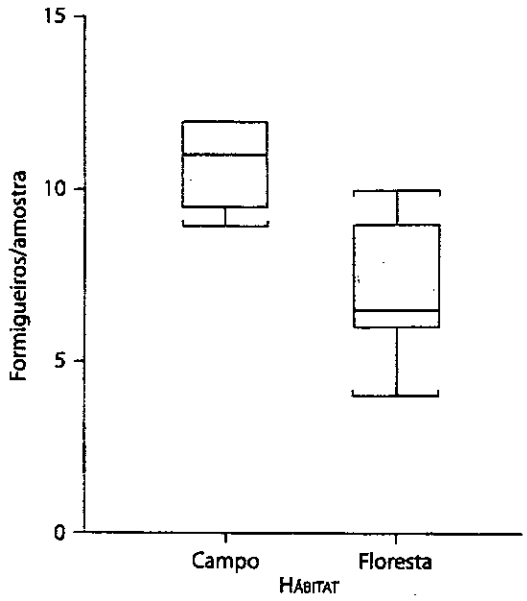


Figura 5.2 *Box plot* dos dados da Tabela 5.1. A linha central dentro da caixa é a mediana dos dados. A caixa inclui 50% dos dados. O topo indica o 75° percentil (quartil superior) dos dados, e o fundo, o 25° percentil (quartil inferior) dos dados. As linhas verticais se estendem até o decil superior e inferior (90° e 10° percentil). Para as amostras de campo, existem tão poucos dados que o 75° e o 90° percentis não diferem. Quando os dados possuem distribuição assimétrica ou discrepâncias (*outliers*), os *box plots* podem ser mais informativos que o gráfico de barras padrão, como o da Figura 5.1.

ANÁLISES DE MONTE CARLO

As análises de Monte Carlo envolvem um número de métodos no qual os dados são aleatorizados* ou misturados de forma que as observações sejam rearranjadas a diferentes tratamentos ou grupos. Essa randomização[2] especifica a hipótese nula sob consideração: de que o padrão nos dados não é diferente daquele que esperaríamos se as observações tivessem sido atribuídas de forma aleatória aos diferentes grupos. Existem quatro passos na análise de Monte Carlo:

[2] Alguns estatísticos distinguem os métodos de Monte Carlo, em que as amostras são tiradas de uma distribuição estatística conhecida ou especificada, dos **testes de randomização**, nos quais os dados existentes são embaralhados, porém nenhum pressuposto é feito acerca da distribuição subjacente. Neste livro, usamos os métodos de Monte Carlo para indicar testes de randomização. Outro conjunto de métodos inclui o **bootstrap**, em que as estatísticas são estimadas pela sub-amostrando repetidamente, com reposição, um conjunto de dados. Ainda, outro conjunto de métodos inclui o **jackknife**, em que a variabilidade em um conjunto de dados é estimada excluindo sistematicamente cada observação e recalculando a estatística (*ver* Figura 9.8, e a sessão "Análises Discriminantes" no Capítulo 12). Monte Carlo, é claro, é a famosa cidade turística de apostas da Riviera, cujos cidadãos não pagam impostos e são proibidos de entrar nas salas de jogos.

* N. deT. Randomização vem da palavra inglesa *Random*, de origem francesa, usada na expressão *at random*, cujo sentido é "ao acaso", "a esmo", "sem seleção ou critério de escolha". O equivalente vernáculo do verbo *randomizar* poderia ser *casualizar*, *acidentalizar* ou *aleatorizar*.

1. Especificar uma estatística-teste ou índice para descrever o padrão dos dados.
2. Criar uma distribuição da estatística-teste que seria esperada sob a hipótese nula.
3. Decidir sobre um teste unicaudal ou bicaudal.
4. Comparar a estatística-teste observada a uma distribuição de valores simulados e estimar um valor de *P* apropriado como probabilidade de cauda (conforme descrito no Capítulo 3).

Passo 1: especificando o teste estatístico

Para essa análise empregaremos como medida do padrão a diferença absoluta entre as médias das amostras de florestas e de campos (nomes dos rótulos), ou DIFF:

$$DIF_{obs} = |10{,}75 - 7{,}00| = 3{,}75$$

O subscrito "obs" indica que o valor de DIF é calculado para os dados observados. A hipótese nula é que a DIF_{obs} igual a 3,75 está dentro do que seria esperado pela amostragem aleatória. A hipótese alternativa poderia ser que a DIF_{obs} igual a 3,75 é maior do que seria esperado ao acaso.

Passo 2: criando a distribuição nula

A seguir, estime qual seria a DIF se a hipótese alternativa fosse verdadeira. Para isso, use o computador (ou as cartas de um baralho) para de maneira aleatória rearranjar os rótulos floresta e campo aos dados. Com a planilha aleatorizada ainda existem 4 amostras de campo e 6 de floresta, mas seus rótulos (Campo ou Floresta) serão aleatoriamente rearranjados (Tabela 5.3). Note que, na planilha de dados misturados, muitas observações são colocadas no mesmo grupo, como nos dados originais. Isso acontecerá com mais frequência em conjuntos de dados pequenos. Depois, calcule

TABELA 5.3 Randomização de Monte Carlo dos rótulos dos hábitats da Tabela 5.1

Hábitat	Número de formigueiros por parcela
Campo	9
Campo	6
Floresta	4
Floresta	6
Campo	7
Floresta	10
Floresta	12
Campo	9
Floresta	12
Floresta	10

Na análise de Monte Carlo, os rótulos das amostras da Tabela 5.1 são rearranjados aleatoriamente entre as diferentes amostras. Após, a diferença entre as médias dos dois grupos, DIF, é registrada. Esse procedimento é repetido muitas vezes, gerando uma distribuição de valores de DIF (Figura 5.3).

a estatística das amostras para o conjunto de dados aleatorizado (Tabela 5.4). Para esse conjunto de dados, $DIF_{sim} = |7,75 - 9,00| = 1,25$. O subscrito "sim" indica que o valor de DIF foi calculado para os dados aleatorizados ou simulados.

No primeiro conjunto de dados simulado, a diferença entre a média dos dois grupos ($DIF_{sim} = 1,25$) é menor que a observada com os dados reais ($DIF_{obs} = 3,75$). Tal resultado sugere que as médias das amostras de floresta e de campo podem diferir mais que o esperado sob a hipótese nula de atribuição aleatória.

Contudo, existem diversas e numerosas combinações aleatórias que podem ser produzidas rearranjando os rótulos das amostras.[3] Alguns têm valores de DIF_{sim} relativamente maiores e outros têm valores de DIF_{sim} menores. Repita esse exercício de rearranjar muitas vezes (em geral, 1.000 ou mais), e então ilustre a distribuição dos valores de DIF com um histograma (Figura 5.3) e as estatísticas descritivas (Tabela 5.5).

A DIF média do conjunto de dados simulados foi de somente 1,46, comparada ao valor observado de 3,75 para os dados originais. O desvio-padrão de 2,07 poderia ser utilizado para construir um intervalo de confiança (*ver* Capítulo 3), mas não deve ser usado, pois a distribuição dos valores simulados (Figura 5.3) tem uma longa cauda à direita e não segue uma distribuição normal. Um intervalo aproximado de 95% pode ser derivado a partir dessa análise de Monte Carlo, identificando os percentis de 2,5, superior e inferior, dos dados do conjunto de valores de DIF simulados, e usar esses valores como limite superior e inferior do intervalo de confiança (Efron, 1982).

Passo 3: decidindo sobre um teste uni ou bicaudal

A seguir, decida se vai usar um **teste unicaudal** ou um **teste bicaudal.** A "cauda" se refere às áreas localizadas à extrema esquerda ou direita sob a função de densidade de probabilidade (*ver* Figura 3.7). São essas caudas da distribuição as utilizadas para

TABELA 5.4 Estatística descritiva para os dados aleatorizados da Tabela 5.3

Hábitat	*N*	Média	Desvio-padrão
Floresta	6	9,00	3,286
Campo	4	7,75	1,500

Na análise de Monte Carlo, esses valores representam a média e o desvio-padrão em cada grupo após um simples rearranjo dos rótulos. A diferença entre as duas médias (DIF = |7,75 – 9,00| = 1,25) é a estatística-teste.

[3] Com 10 amostras divididas em um grupo de 6 e um grupo de 4, existem

$$\binom{10}{4} = 210$$

combinações que podem ser criadas embaralhando os rótulos (*ver* Capítulo 2). No entanto, o número de valores únicos de DIF possíveis é um pouco menor, pois algumas das amostras têm contagens de formigueiros idênticas.

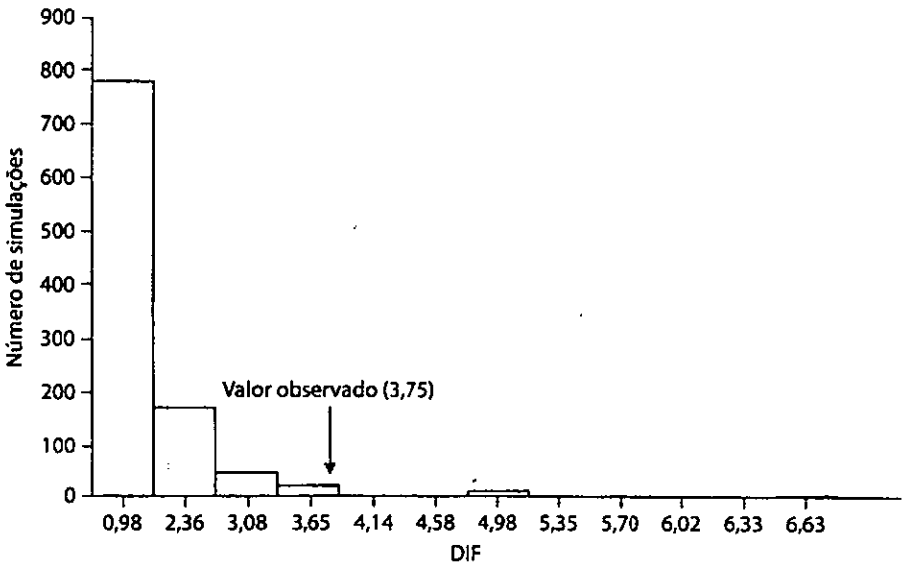


Figura 5.3 Análise de Monte Carlo dos dados da Tabela 5.1. Para cada randomização, os rótulos (Floresta/Campo) foram aleatoriamente rearranjados entre as réplicas. Depois, a diferença (DIF) entre as médias dos dois grupos foi calculada. Este histograma ilustra a distribuição de valores de DIF de 1.000 randomizações. A seta indica o único valor da DIF observada no conjunto de dados real (3,75). Essa DIF se assenta bem na cauda direita da distribuição. A DIF observada de 3,75 foi sempre maior que ou igual a todos, com exceção de 36 valores de DIF simulados. Portanto, sob a hipótese nula de arranjo aleatório das amostras aos grupos, a probabilidade de encontrar essa observação (ou uma mais extrema) é de 36/1.000 = 0,036.

determinar o ponto de corte de um teste estatístico a um $P = 0,05$ (ou qualquer outro nível de probabilidade). Um teste unicaudal usa apenas uma cauda da distribuição para estimar o valor de *P*. Um teste bicaudal usa as duas caudas da distribuição para estimar o valor de *P*. No primeiro caso, os 5% da área sob a curva estão localizados em uma das caudas da distribuição. No segundo, cada cauda da distribuição deve englobar 2,5% da área. Assim, um teste bicaudal requer valores mais extremos para atingir a significância estatística do que um teste unicaudal.

Contudo, o valor de corte não é o tópico mais importante ao decidir quando usar um teste com uma ou com duas caudas. Mais importante é a natureza da

TABELA 5.5 Estatísticas descritivas para os 1.000 valores simulados de DIF da Figura 5.3

Variável	*N*	Média	Desvio-Padrão
DIF_{sim}	1.000	1,46	2,07

Na análise de Monte Carlo, cada um desses valores foi criado rearranjando os rótulos dos dados da Tabela 5.1 e calculando a diferença entre as médias dos dois grupos (DIF). O cálculo do valor de *P* não requer que a DIF siga uma distribuição normal, pois o valor de *P* é determinado diretamente pela localização da DIF observada no histograma (*ver* Tabela 5.6).

variável resposta e a hipótese que está sendo testada. Neste caso, a variável resposta era a DIF, a diferença *absoluta* nas médias das amostras de floresta e campos. Para a variável DIF, um teste unicaudal é mais apropriado para valores de DIF excepcionalmente grandes. Por que não usar um teste bicaudal para DIF? A cauda inferior da distribuição de DIF_{sim} poderia representar os valores de DIF que são excepcionalmente pequenos quando comparados à hipótese nula. Em outras palavras, um teste com duas caudas poderia testar também a possibilidade das médias de florestas e campos serem mais similares que o esperado ao acaso. Esse teste para a similaridade extrema não é biologicamente informativo, então a atenção é restrita à cauda superior da distribuição, a qual representa os casos em que a DIF é excepcionalmente grande quando comparada à distribuição nula.

Como a análise poderia ser modificada para usar um teste bicaudal? Em vez de usar o valor absoluto de DIF, use a diferença média entre amostras de floresta e campo (DIF*). Diferente da DIF, a DIF* pode assumir valores tanto positivos quanto negativos. A DIF* será positiva se a média do campo for maior que a da floresta. A DIF* será negativa se a média de campo for menor que a da floresta. Como antes, randomize os dados e crie a distribuição de DIF^*_{sim}. Neste caso, contudo, um teste bicaudal poderia detectar casos em que a média de campo seria excepcionalmente grande se comparada à da floresta (DIF* positiva) e casos em que a média de campo seria excepcionalmente pequena comparada à da floresta (DIF* negativa).

Quando estiver usando a análise de Monte Carlo, paramétrica ou bayesiana, você deveria estudar com cuidado a variável resposta que está usando. Como poderia interpretar um valor extremamente grande ou pequeno da variável resposta relativo à hipótese nula? A resposta a essa questão irá ajudá-lo a decidir quando um teste com uma ou com duas caudas é mais apropriado.

Passo 4: calculando a probabilidade de cauda

O passo final é estimar a probabilidade de obter DIF_{obs} ou um valor mais extremo, visto que a hipótese nula é verdadeira [$P(\text{dados}|H_0)$]. Para fazer isso, examine o conjunto de valores simulados de DIF (histograma da Figura 5.3) e conte quantas vezes a DIF_{obs} foi maior que, igual a, ou menor que cada um dos valores de DIF_{sim}.

Em 29 de 1.000 randomizações, $DIF_{sim} = DIF_{obs}$, então a probabilidade de obter a DIF_{obs} sob a hipótese nula é de 20/1.000 = 0,029 (Tabela 5.6). Contudo, quando calculamos uma estatística-teste, em geral não estamos tão interessados na **probabilidade exata** quanto na **probabilidade em cauda**. Ou seja, queremos saber as chances de obter uma observação *grande como* ou *maior que* os dados reais, desde que a hipótese nula seja verdadeira. Em 7 de 1.000 randomizações, $DIF_{sim} > DIF_{obs}$. Deste modo, a probabilidade de que $DIF_{obs} \geq 3{,}75$ é (7 + 29) / 1.000 = 0,036. Esta probabilidade em cauda é a frequência de obter o valor observado (29/1.000) mais a de obter um resultado mais extremo (7/1.000).

TABELA 5.6 Cálculo da probabilidade em cauda na análise de Monte Carlo

Desigualdade	*N*
$DIF_{sim} > DIF_{obs}$	7
$DIF_{sim} = DIF_{obs}$	29
$DIF_{sim} < DIF_{obs}$	964

Comparações da DIF_{obs} (diferença absoluta entre a média dos dois grupos com os dados originais) com as DIF_{sim} (diferença absoluta na média dos dois grupos após a randomização das atribuições dos grupos). *N* é o número de simulações (de 1.000) para os quais a desigualdade foi obtida. Como a $DIF_{sim} \geq DIF_{obs}$ em 7+29 = 36 simulações de 1.000, a probabilidade em cauda sob a hipótese nula de encontrar uma DIF_{obs} neste extremo é de 36/1.000 = 0,036.

Siga os procedimentos e as interpretações dos valores de *P* que discutimos no Capítulo 4. Com uma probabilidade em cauda de 0,036, esses dados não teriam ocorrido caso a hipótese nula fosse verdadeira.

Premissas do método de Monte Carlo

Os métodos de Monte Carlo recaem sobre três premissas:

1. Os dados coletados representam amostras aleatórias e independentes.
2. A estatística-teste descreve o padrão de interesse.
3. A randomização cria uma distribuição nula apropriada para a questão.

As premissas 1 e 2 são comuns a todas as análises estatísticas. A 1 é a mais crítica, mas é também a mais difícil de se confirmar, como discutiremos no Capítulo 6. A 3 é fácil de se satisfazer neste caso. A estrutura amostral e a hipótese nula sendo testadas são muito simples. Para questões mais complexas, contudo, o método de randomização apropriado pode não ser óbvio e, sim, haver mais de uma forma de construir a distribuição nula (Gotelli & Graves, 1996).

Vantagens e desvantagens do método de Monte Carlo

A principal vantagem do método de Monte Carlo é que ele deixa claras e explícitas as premissas subjacentes e a estrutura da hipótese nula. Em contraste, análises paramétricas convencionais frequentemente ofuscam essas funcionalidades, talvez porque os métodos sejam muito familiares. Outra vantagem do método de Monte Carlo sobre as análises paramétricas é que ele não requer a premissa de que os dados foram amostrados de uma distribuição específica, como a normal. Por fim, as simulações de Monte Carlo permitem que você personalize seu teste estatístico para questões e conjuntos de dados em particular, em vez de ter que forçá-los em um teste convencional que pode não ser o mais eficaz para sua questão, ou cujas premissas podem não corresponder ao delineamento amostral de seus dados.

A principal desvantagem do método de Monte Carlo é que ele requer computação intensiva e não é incluído na maioria dos pacotes estatísticos tradicionais (ver, porém, Gotelli & Entsminger, 2003). Conforme os computadores se tornam mais rápidos, antigas limitações dos métodos computacionais estão desaparecendo e não há razão para que mesmo análises estatísticas muito complexas não possam ser executadas com uma simulação de Monte Carlo. No entanto, até que tais rotinas se tornem amplamente disponíveis, os métodos de Monte Carlo estão disponíveis apenas para aqueles que sabem uma linguagem de programação e podem escrever seus próprios programas.[4]

Uma segunda desvantagem das análises de Monte Carlo é psicológica. Alguns cientistas são receosos acerca desses métodos porque diferentes análises do mesmo conjunto de dados podem produzir respostas um pouco diferentes. Por exemplo, nós re-executamos a análise sobre os dados da Tabela 5.1 dez vezes e obtivemos valores de *P* que variaram de 0,030 até 0,046. A maioria dos pesquisadores prefere as análises paramétricas, que são mais objetivas e resultam no mesmo valor de *P* cada vez que a análise é repetida.

A última desvantagem é que o domínio de uma inferência para uma análise de Monte Carlo é subitamente mais restritivo que para uma do tipo paramétrica, que assume uma distribuição especificada e permite inferências acerca da população de origem subjacente, a partir da qual os dados foram amostrados. Estritamente falando, as inferências dos métodos de Monte Carlo (pelo menos aquelas com base em simples testes de randomização) são limitadas aos dados específicos coletados. Contudo, se uma amostra é representativa da população de origem, os resultados podem ser cuidadosamente generalizados.

[4] Uma das coisas mais importantes que você pode fazer é reservar um tempo para aprender uma linguagem de programação de verdade. Apesar de alguns indivíduos serem proficientes na programação de macros em planilhas, os macros são práticos apenas para os cálculos mais elementares; para coisas mais complexas, é mais simples escrever algumas linhas de códigos de computador do que lidas com os passos convolutos (e propensos a erro) necessários para escrever macros. Na atualidade existem muitas linguagens de computador para você escolher entre elas. O Gotelli prefere Delphi, que é uma versão de alto nível de Pascal com uma interface de usuário gráfica; o Ellison prefere S-Plus ou sua versão livre, R, que inclui muitas funções matemáticas e estatísticas. Infelizmente, aprender a programar é como aprender uma língua estrangeira – leva tempo e prática, e não há uma recompensa imediata. É lamentável, mas nossa cultura acadêmica não encoraja o aprendizado de habilidades de programação (ou outras línguas que não o português). Mas se você quiser superar a íngreme curva do aprendizado, a recompensa científica é grande. A melhor forma de começar não é tomar aulas, mas adquirir um *software* de linguagem, trabalhar ao longo dos exemplos do manual, e então começar com um problema que o interesse. O livro *The Ecological Detective* (1997), de Hilborn e Mengel, contém uma excelente série de exercícios ecológicos para exercitar suas habilidades de programação. Programar não somente irá livrá-lo da cadeia dos *softwares* enlatados, mas também aguçar suas habilidades analíticas e lhe dar novos *insights* em modelos ecológicos e estatísticos. Você terá um entendimento profundo de um modelo ou de uma estatística, desde que os tenha programado com sucesso.

ANÁLISES PARAMÉTRICAS

As análises paramétricas referem-se ao grande corpo de teorias e testes estatísticos construídos sob a premissa de que os dados analisados foram amostrados de uma distribuição específica. A maioria dos testes estatísticos familiares aos ecólogos e cientistas da área ambiental especifica a distribuição normal. Os parâmetros da distribuição (p. ex., as médias da população μ e a variância σ^2) são estimados e usados para calcular a probabilidade de cauda para uma hipótese nula verdadeira. Uma grande estrutura estatística tem sido construída acerca da premissa simplista da normalidade dos dados. Praticamente 80% a 90% do que é ensinado nos textos-padrão de estatística recaem sob esse tema. Aqui, usamos um método paramétrico comum, a **análise de variância** (ANOVA) para testar diferenças entre a média dos grupos para os dados amostrais.[5] Existem três passos na análise paramétrica:

1. Especifique a estatística-teste.
2. Especifique a distribuição nula.
3. Calcule a probabilidade de cauda.

Passo 1: especificando a estatística-teste

A análise de variância paramétrica assume que os dados foram tirados de uma distribuição normal, ou gaussiana. A média e a variância dessas curvas podem ser estimadas a partir dos dados amostrais (Tabela 5.1), usando as Equações 3.1 e 3.9. A Figura 5.4 mostra as distribuições usadas na análise de variância paramétrica. Os dados originais são arranjados no eixo-x e cada cor representa um hábitat diferente (círculos pretos para as amostras da floresta e círculos verdes para as do campo).

Primeiro, considere a hipótese nula: que ambos os conjuntos de dados foram tirados de uma distribuição normal subjacente única, estimada a partir da média

Sir Ronald Fisher

[5] O arcabouço da teoria estatística paramétrica moderna é amplamente desenvolvida pelo notável Sir Ronald Fisher (1890-1962), e a razão-F é nomeada em sua honra (embora o próprio Fisher ache que essa razão precisava de mais estudos e refinamentos). Fisher recebeu a cadeira de honra de Genética na Cambridge University, de 1943 a 1957, e fez contribuições fundamentais para as teorias de genética de populações e evolução. Na estatística, ele desenvolveu a análise de variância (ANOVA) para analisar a produção de cultivos em sistemas de agricultura, em que pode ser difícil ou impossível replicar os tratamentos. Hoje, muitas das mesmas restrições desafiam os ecólogos no delineamento de seus experimentos, por isso os métodos de Fisher continuam tão úteis. Seu livro clássico, *The Design of Experiments* (1935), ainda é uma leitura agradável e que vale a pena. É irônico pensar que Fisher tenha ficado preocupado acerca de seus próprios métodos ao lidar com experimentos que ele podia delinear bem, enquanto hoje muitos ecólogos aplicam os métodos dele para avaliar observações de experimentos naturais maldelineados (*ver* Capítulos 4 e 6). (A fotografia é cortesia do Memorial de Ronald Fisher.)

e da variância de todos os dados (linha pontilhada; média = 8,5, desvio-padrão = 2,54). A hipótese alternativa é de que as amostras foram tiradas de duas populações diferentes, cada uma delas pode ser caracterizada por uma distribuição normal diferente, uma para floresta e outra para campo. Cada distribuição tem sua própria média, embora assumamos que a variância é a mesma (ou similar) para cada um dos grupos. Essas duas curvas também são ilustradas na Figura 5.4, usando as estatísticas descritivas calculadas na Tabela 5.2.

Como a hipótese nula é testada? Quanto mais próximas as curvas para os dados de floresta e campo, mais provável é que os dados tenham sido coletados visto que a hipótese nula é verdadeira, e a linha pontilhada melhor os representa. Inversamente, quanto mais separadas forem as curvas, mais improvável é que os dados representem uma única população como média e variância comuns. A área sobreposta entre as duas distribuições (sombreada, na Figura 5.4) deveria ser uma medida de quão próximo ou quão distante as distribuições estão.

A contribuição de Fisher foi de quantificar essa sobreposição como a razão das duas variáveis. A primeira é a quantidade de variação entre grupos, que po-

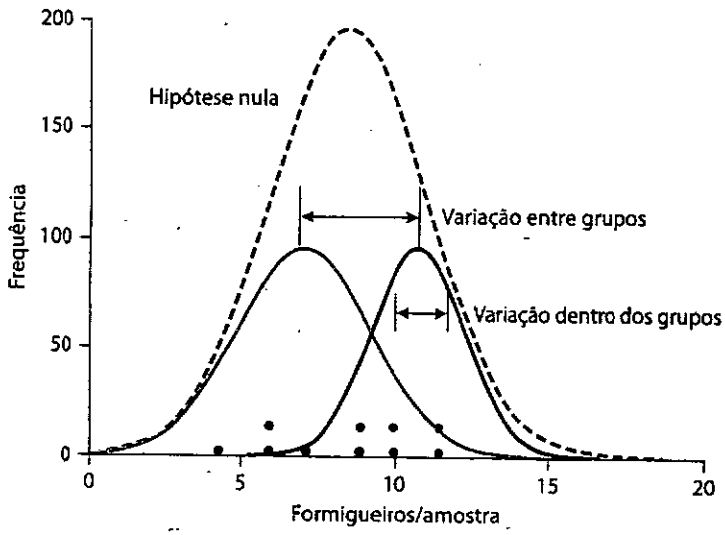


Figura 5.4 Distribuições normais fundamentadas nos dados amostrais da Tabela 5.1. Os dados da Tabela 5.1 são mostrados como símbolos (círculos pretos, amostras de floresta; círculos verdes, amostras de campo) indicativos do número de formigueiros registrados em cada parcela. A hipótese nula é de que todos os dados foram tirados da mesma população, cuja distribuição normal é mostrada com a linha pontilhada. A hipótese alternativa é de que cada hábitat tem sua própria média (e variância) distintas, indicadas pelas duas distribuições normais menores. Quanto menor a sobreposição sombreada das duas distribuições, mais improvável que a hipótese nula seja verdadeira. As medidas de variação entre e dentro dos grupos são usadas para calcular a razão-F e testar a hipótese nula.

demos pensar como a variância (ou o desvio-padrão) das médias dos dois grupos. A segunda é a quantidade de variação dentro de cada grupo, que podemos pensar como a variância das observações em torno de suas respectivas médias. A **razão-F de Fisher** pode ser interpretada como a existente entre essas duas fontes de variação:

F= (variância entre grupos + variância dentro dos grupos) / variância dentro dos grupos (5.1)

No Capítulo 10, explicaremos em detalhes como calcular o numerador e o denominador da razão-F. Por enquanto, apenas enfatizaremos que a razão mede o tamanho relativo das duas fontes de variação nos dados: variação entre grupos e variação dentro dos grupos. Para esses dados, a razão-F é calculada como 33,75/3,84 = 8,78.

Em uma ANOVA, a razão-F é a estatística-teste que descreve (com um único número) o padrão das diferenças entre as médias dos diferentes grupos sob comparação.

Passo 2: especificando a distribuição nula

A hipótese nula é de que todos os dados foram tirados da mesma população, de forma que quaisquer diferenças entre as médias dos dois grupos não sejam maiores que esperado ao acaso. Se essa hipótese nula for verdadeira, então a variação entre grupos será pequena, e esperamos encontrar uma razão-F de 1,0. A razão-F será correspondentemente maior que 1,0 se as médias dos grupos estão muito separadas (grande variação entre grupos) em relação à variação dentro dos grupos.[6] Neste exemplo, a razão-F de 8,78 é quase 10 vezes maior que o valor esperado de 1,0, um resultado que parece improvável caso a hipótese nula fosse verdadeira.

Passo 3: calculando a probabilidade de cauda

O valor de *P* é uma estimativa da probabilidade de obter uma razão-F ≥ 8,78, dado que a hipótese nula fosse verdadeira. A Figura 5.5 mostra a distribuição teórica da razão-F e a razão-F observada de 8,78, que está no extremo da cauda direita da distribuição. Qual é sua probabilidade de cauda (ou valor de *P*)?

O valor de *P* dessa razão-F é calculado como a massa de probabilidade da **distribuição da razão-F** (ou área sob a curva) igual ou maior que a razão-F observada. Para esses dados, a probabilidade de obter uma razão-F tão grande ou maior que 8,78 (dados dois grupos e um N total de 10) é igual a 0,018. Como o valor de

[6] Como na análise de Monte Carlo, existe uma possibilidade teórica de obter uma razão-F que é menor que o esperado ao acaso. Assim, as médias dos dois grupos são inusitadamente similares e diferem menos que o esperado, se a hipótese nula fosse verdadeira. As razões-F inusitadamente pequenas raras vezes são vistas na literatura ecológica, embora Schluter (1990) as tenha usado como um índice de concordância do tamanho corporal entre espécies e para convergência da comunidade.

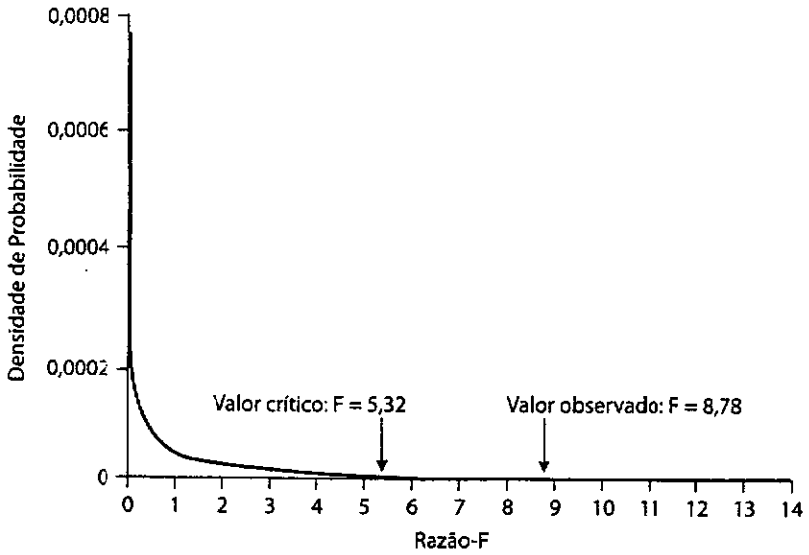


Figura 5.5 Distribuição-F teórica. Quanto maior a razão-F observada, mais improvável ela seria caso a hipótese nula fosse verdadeira. O valor crítico para essa distribuição é igual a 5,32; a área sob a curva além deste ponto é igual a 5% de toda área sob a curva. A razão-F observada de 8,78 está além do valor crítico, assim *P*(dados das formigas | hipótese nula) ≤ 0,05. De fato, a verdadeira probabilidade *P*(dados das formigas | hipótese nula) = 0,018, pois a área sob a curva à direita da razão-F observada representa 1,8% da área total sob a curva. Compare esse resultado ao valor de *P* de 0,036 da análise de Monte Carlo (Figura 5.3).

P é menor que 0,05, consideramos improvável que uma razão-F tão alta ocorra apenas por acaso, e rejeitamos a hipótese nula de que nossos dados tenham sido amostrados de uma única população.

Premissas do método paramétrico

Existem duas premissas para todas as análises paramétricas:

1. Os dados coletados representam amostras aleatórias e independentes.
2. Os dados foram amostrados de uma distribuição específica.

Como vimos, a primeira premissa de amostragem aleatória e independente para as análises de Monte Carlo é sempre a mais importante em qualquer outra. A segunda premissa normalmente é satisfeita, porque distribuições normais (forma de sino) são ubíquas e ocorrem com frequência no mundo real.

Testes paramétricos específicos normalmente incluem premissas adicionais. Por exemplo, a ANOVA ainda assume que a variância dentro de cada grupo sob comparação é igual (*ver* Capítulo 10). Se o tamanho amostral é grande, tal premissa pode ser modestamente violada e os resultados ainda assim serão robustos.

Porém, se o tamanho amostral for pequeno (como neste exemplo) essa premissa é mais importante.

Vantagens e desvantagens do método paramétrico

A vantagem do método paramétrico é que ele utiliza um arcabouço poderoso com base em distribuições de probabilidade conhecidas. A análise que apresentamos foi muito simples, mas existem muitos testes paramétricos apropriados para delineamentos experimentais e amostrais complexos (*ver* Capítulo 7).

Embora a análise paramétrica seja intimamente associada ao teste da hipótese nula estatística, ela pode não ser tão poderosa quanto os sofisticados modelos de Monte Carlo, que são adaptados a questões ou dados particulares. Em contraste à análise bayesiana, a paramétrica raras vezes incorpora informação *a priori* ou resultados de outros experimentos. As análises bayesianas serão alvo do próximo tópico – após um breve desvio pelas estatísticas não paramétricas.

Análises não paramétricas: um caso especial de análise de Monte Carlo

As estatísticas não paramétricas têm base na análise de dados ranqueados. No exemplo das formigas, poderíamos ranquear as observações da maior para a menor e, então, calcular estatísticas com base nas somas, nas distribuições ou em outras medidas sintéticas desses ranques. Análises não paramétricas não assumem uma distribuição paramétrica específica (daí o nome), mas elas ainda requerem amostras independentes e aleatórias (como todas as análises estatísticas). Um teste não paramétrico, em essência, é uma análise de Monte Carlo de dados ranqueados, e tabelas estatísticas não paramétricas fornecem valores de P que poderiam ser obtidos por testes de randomização de ranques. Portanto, já descrevemos a lógica e os procedimentos gerais para esses testes.

Embora elas em geral sejam usadas por ecólogos e cientistas da área ambiental, não somos a favor das análises não paramétricas por três razões. Primeiro, o uso de dados ranqueados desperdiça informações que estão presentes nas observações originais. Uma análise de Monte Carlo dos dados brutos é bem mais informativa e, com frequência, mais poderosa. Uma justificativa que é dada para o uso de uma análise não paramétrica é que os dados ranqueados podem ser mais robustos aos erros de medidas. Contudo, se as observações originais são tão propensas a erro que somente os ranques sejam confiáveis, provavelmente uma boa ideia seja refazer o experimento usando métodos de mensuração que ofereçam melhor acurácia. Segundo, relaxar com a premissa de uma distribuição paramétrica (p. ex., normalidade) não é uma grande vantagem, pois as análises paramétricas muitas vezes são robustas às violações dessa premissa (graças ao Teorema do Limite Central). Terceiro, os métodos não paramétricos são disponíveis apenas para delineamentos experimentais muito simples e não podem incorporar com facilidade covariáveis ou estruturas em bloco. Temos

observado que praticamente todas as necessidades de análises estatísticas de dados ecológicos e ambientais podem ser satisfeitas por abordagens paramétricas, de Monte Carlo e bayesianas.

ANÁLISES BAYESIANAS

A análise bayesiana é a terceira maior estrutura para análises de dados. Os cientistas com frequência acreditam que seus métodos são "objetivos", pois tratam cada experimento como uma *tábula rasa*: a simples hipótese nula estatística de variação aleatória reflete a ignorância acerca da causa e do efeito. Em nosso exemplo da densidade de formigueiros em florestas e campos, nossa hipótese nula era a de que os dois são iguais, ou de que o fato de estar na floresta ou no campo não possui nenhum efeito consistente da densidade de formigueiros. Apesar de ser possível que ninguém nunca tenha investigado isso, trata-se de algo extremamente improvável; nossa revisão da literatura sobre biologia de formigas que nos levou a realizar este estudo específico. Então, por que não usar dados que já existem para formular nossas hipóteses? Se nosso único objetivo é o hipotético-dedutivo de falsear uma hipótese nula, e todos os dados prévios indicam que florestas e campos diferem na densidade de formigueiros, é muito provável que nós, também, falsearemos a hipótese nula. Dessa forma, não precisaríamos gastar tempo e energia fazendo o estudo novamente.

Os bayesianos argumentam que poderíamos progredir mais especificando a diferença observada (p. ex., expressada como a DIF ou a estatística-F descrita nas sessões anteriores), e então usar nossos dados para ampliar os resultados anteriores de outros pesquisadores. A análise bayesiana nos permite fazer exatamente isso, bem como quantificar a probabilidade da diferença observada. Eis a diferença mais importante entre os métodos bayesiano e frequentista.

Existem seis passos na inferência bayesiana:

1. Especificar a hipótese.
2. Especificar os parâmetros como variáveis aleatórias.
3. Especificar a probabilidade *a priori*.
4. Calcular a verossimilhança.
5. Calcular a distribuição de probabilidades *a posteriori*.
6. Interpretar os resultados.

Passo 1: especificando a hipótese

O objetivo primário da análise bayesiana é determinar a probabilidade da hipótese *aplicando* os dados que foram coletados: $P(\text{H}|\text{dados})$. A hipótese precisa ser bastante específica, e quantitativa. Em nossa análise paramétrica da densidade de formigueiros, a hipótese de interesse (i. e., a hipótese alternativa) foi que as

amostras foram tiradas de duas populações com médias diferentes e mesma variância, uma para floresta e outra para campo. Não testamos essa hipótese diretamente. Em vez disso, testamos a hipótese nula: o valor observado do F não era maior que o esperado caso as amostras fossem tiradas de uma única população. Encontramos que a razão-F observada era bastante improvável ($P = 0{,}018$) e rejeitamos a hipótese nula.

Poderíamos especificar com mais precisão a hipótese nula e a alternativa acerca do valor da razão-F. Antes de podermos especificar essas hipóteses, precisamos saber o valor crítico para distribuição-F na Figura 5.5. Em outras palavras, quão grande um valor de F precisa ser para ter um valor de $P \leq 0{,}05$?

Para 10 observações de formigas em dois grupos (campo e floresta), o valor crítico da distribuição-F (i. e., o valor para o qual a área sob a curva é igual a 5% da área total) é igual a 5,32 (*ver* Figura 5.5). Assim, qualquer razão-F observada que seja maior ou igual a 5,32 seria motivo para rejeitar a hipótese nula. Lembre-se de que o método hipotético-dedutivo geral declara que a probabilidade da hipótese nula é $P(\text{dados}|H_0)$. No exemplo dos formigueiros, os dados resultaram em uma razão-F igual a 8,78. Se a hipótese nula for verdadeira, a razão-F observada deveria ser uma amostra aleatória da distribuição F mostrada na Figura 5.5. Por essa razão, questionamos o que é $P(F_{obs} = 8{,}78 \mid F_{teórica})$?

Em contraste, a análise bayesiana procede invertendo essa declaração de probabilidade: qual é a probabilidade da hipótese *aplicando* os dados que coletamos [$P(H \mid \text{dados})$]? Os dados dos formigueiros podem ser expressos como F = 8,78. Como as hipóteses seriam expressas em termos da distribuição-F? A hipótese nula é de que os formigueiros foram amostrados de uma única população. Neste caso, o valor esperado para a razão-F é pequeno (F < 5,32, o valor crítico). A hipótese alternativa é de que os formigueiros foram amostrados de duas populações, na qual a razão-F seria grande (F = 5,32). Portanto, a análise bayesiana da hipótese alternativa calcula $P(F \geq 5{,}32 \mid F_{obs} = 8{,}78)$. Pelo Primeiro Axioma da Probabilidade, $P(F < 5{,}32 \mid F_{obs}) = 1 - P(F \geq 5{,}32 \mid F_{obs})$.

Uma modificação do Teorema de Bayes (apresentada no Capítulo 1) nos permite calcular diretamente $P(\text{hipótese} \mid \text{dados})$:

$$P(\text{hipótese} \mid \text{dados}) = \frac{P(\text{hipótese})\, P(\text{dados} \mid \text{hipótese})}{P(\text{dados})} \qquad (5.2)$$

Na Equação 5.2, $P(\text{hipótese} \mid \text{dados})$ no lado esquerdo da equação é chamada de **distribuição de probabilidade *a posteriori*** (ou simplesmente os posteriores), e é a quantidade de interesse. O lado direito da equação consiste em uma fração. No numerador, o termo $P(\text{hipótese})$ é tratado como **distribuição de probabilidades *a priori*** (ou simplesmente os priores), e é a probabilidade da hipótese de interes-

se *antes* de conduzir o experimento. O próximo termo no numerador, *P*(dados | hipótese), é tratado como a **verossimilhança** dos dados; ele reflete a probabilidade de observar os dados conforme a hipótese.[7] O denominador, *P*(dados), é uma constante de normalização que reflete a probabilidade dos dados de acordo com todas as hipóteses possíveis.[8] Como ela é simplesmente uma constante de normalização (e coloca a probabilidade *a posteriori* em uma escala que varia entre [0,1]),

$$P(\text{hipótese} \mid \text{dados}) \propto P(\text{hipótese})\, P(\text{dados} \mid \text{hipótese})$$

(onde $\propto$ significa "é proporcional a") e podemos focar nossa atenção ao numerador.

Retornando às formigas e às razões-F, detemo-nos em $P(F \geq 5{,}32 \mid F_{obs} = 8{,}78)$. Agora já especificamos quantitativamente nossa hipótese nos termos da relação entre a razão-F que observamos (os dados) e o valor crítico de $F \geq 5{,}32$ (a hipótese). Trata-se de uma hipótese mais precisa do que a discutida nas duas sessões anteriores, de que existe uma diferença entre a densidade de formigas em campos e em florestas.

Passo 2: especificando os parâmetros como variáveis aleatórias

Uma segunda diferença fundamental entre as análises frequentista e bayesiana é que, na primeira, os parâmetros (como a verdadeira média das populações $\mu_{floresta}$

[7] Fisher desenvolveu o conceito de verossimilhança (*likelihood*) em resposta ao seu desconforto com os métodos bayesianos de probabilidade inversa:

> *O que tem parecido é que o conceito matemático de probabilidade é inadequado para expressar nossa confiança, ou desconfiança, mental ao fazer tais inferências, e que a quantidade matemática que parece ser apropriada para medir a nossa ordem de preferência entre diferentes populações possíveis de fato não obedece às leis da probabilidade. Para distingui-la da probabilidade, eu tenho usado o termo "verossimilhança" para designar essa quantidade. (Fisher, 1925, p. 10).*

A verossimilhança é representada como *L*(hipótese | dados) e é diretamente proporcional (mas não igual) à probabilidade dos dados *observados* dada a hipótese de interesse: *L*(hipótese | dados) = $cP(\text{dados}_{obs} \mid \text{hipótese})$. Desta forma, a verossimilhança difere do valor de *P* frequentista, pois o valor de *P* expressa a probabilidade das infinitamente possíveis amostras de dados de acordo com a hipótese nula estatística (Edwards, 1992). A verossimilhança é usada extensivamente em abordagens que usam teoria de informação na inferência estatística (p. ex., Hilborn & Mangel, 1997; Burnham & Anderson, 2002), e é uma parte central da inferência bayesiana. Entretanto, a verossimilhança não segue os axiomas da probabilidade. Como a linguagem da probabilidade é uma forma mais consistente de expressar nossa confiança em um determinado resultado, sentimos que as probabilidades de diferentes hipóteses (presentes na escala de 0 a 1) são interpretadas com mais facilidade que as verossimilhanças (ausentes).

[8] O denominador é calculado como:

$$\int P(H_i) P(\text{dados} \mid H_i)\, dH$$

e μ_{campo}, seus desvios-padrão σ^2, ou a razão-F) são *fixos*. Em outras palavras, assumimos que há um *valor verdadeiro* para a densidade de formigueiros em campos e florestas (ou ao menos nos campos e nas florestas que amostramos), e estimamos aqueles parâmetros a partir de nossos dados. Em contraste, a análise bayesiana considera esses parâmetros como sendo *variáveis aleatórias*, com seus próprios parâmetros associados (p. ex., médias e variâncias). Assim, por exemplo, em vez da média da população de colônias de formigas em campo ser um valor fixo μ_{campo}, ela pode ser expressa como uma variável aleatória com sua própria média e variância: $\mu_{campo} \sim N(\lambda_{campo}, \sigma^2)$. Note que a variável aleatória representando as colônias de formigas não precisa ser normal. O tipo de variável aleatória usada para cada parâmetro da população deve refletir a realidade biológica, e não uma conveniência matemática ou estatística. Nesse exemplo, contudo, é razoável descrever a densidade de colônias de formigas como uma variável aleatória normal: $\mu_{campo} \sim N(\lambda_{campo}, \sigma^2)$, $\mu_{floresta} \sim N(\lambda_{floresta}, \sigma^2)$.

Passo 3: especificando a distribuição de probabilidades de priores

Como nossos parâmetros são variáveis aleatórias, eles possuem distribuições de probabilidades associadas. Nossas médias desconhecidas das populações (os termos λ_{campo} e $\lambda_{floresta}$) possuem distribuição normal, associados à média e variância desconhecidas. Para fazer os cálculos requeridos pelo Teorema de Bayes, precisamos especificar a distribuição de probabilidade de priores para esses parâmetros – ou seja, qual é a distribuição de probabilidade para essas variáveis aleatórias *antes* de fazermos o experimento?[9]

Temos duas escolhas básicas para especificar *a priori*. Primeiro, podemos vasculhar e reanalisar dados na literatura, falar com especialistas e chegar a estimativas razoáveis para a densidade de formigueiros em campos e florestas. Como alternativa, podemos usar uma priore não informativa, para a qual inicialmente estimamos a densidade de formigueiros como sendo igual a zero e a variância como sendo muito grande. (Neste exemplo, estipulamos a variância como sendo igual a 100.000.) Usando uma priore não informativa equivale a dizer que nós não temos nenhuma informação *a priori*, e que a média pode

[9] A especificação da priores é uma divisão fundamental entre frequentistas e bayesianos. Para um frequentista, especificar uma priore reflete subjetividade por parte do investigador, e assim o uso de uma priore é considerado não científico. Os bayesianos argúem que especificar uma priore torna explícitas todas as premissas escondidas de um investigador, e por isso é uma abordagem mais honesta e objetiva de fazer ciência. Esse argumento durou por séculos (*ver* revisão em Effron, 1986, e em Berger & Berry, 1988), e foi uma das razões para a marginalização dos bayesianos dentro da comunidade científica. Contudo, o advento das modernas técnicas de computacionais permitiu que os bayesianos trabalhassem com priores não informativas, como as que usamos aqui. Acontece que, com priores não informativas, os resultados bayesianos e frequentistas são muito similares, apesar de sua interpretação final permanecer diferente. Tais avanços levaram a uma recente renascença na estatística bayesiana, embora um *software* simples e fácil de usar ainda não exista.

assumir praticamente qualquer valor com probabilidade aproximadamente igual.[10] É claro, no caso de haver mais informação, você pode ser mais específico com sua priore. A Figura 5.6 ilustra a priore não informativa e uma mais informativa (hipotética).

Similarmente, o desvio-padrão σ para μ_{campo} e $\mu_{floresta}$ também possui uma distribuição de probabilidades *a priori*. A inferência bayesiana em geral especifica uma distribuição gamma inversa[11] para as variâncias; como com os priores da média, usamos priores não informativos para variância. Escrevemos isso simbolicamente como $\sigma^2 \sim GI(1.000, 1.000)$ (lê-se "a variância é distribuída como uma distribuição gamma inversa com parâmetros 1.000 e 1.000"). Também calculamos a *precisão* de nossa estimativa da variância, que simbolizamos como τ, onde $\tau = 1/\sigma^2$. Aqui, τ é uma variável aleatória gamma (o inverso de uma gamma inversa é uma gamma), e escrevemos isso simbolicamente como $\tau \sim \Gamma(0{,}001, 0{,}001)$ (lê-se "tau é distribuído como uma distribuição gamma com parâmetros 0,001, 0,001"). A forma dessa distribuição é ilustrada na Figura 5.7. O uso dessa precisão deve fazer algum sentido intuitivo; como nossa estimativa da variabilidade decresce quando temos mais informações, a precisão de nossa estimativa aumenta. Assim, um valor alto de precisão equivale a uma baixa variância do nosso parâmetro estimado.

[10] Você pode se questionar por que não usamos uma distribuição uniforme, na qual todos os valores têm a mesma probabilidade. A razão é que a distribuição uniforme é uma priore imprópria. Como a integral de uma distribuição uniforme é indefinida, não podemos usá-la para calcular a distribuição *a posteriori* do Teorema de Bayes. A priore não informativa N(0,100.000) é praticamente uniforme sobre uma grande parte, mas ela pode ser integrada. *Ver* Carlin e Louis (2000) para uma discussão adicional sobre priores impróprios e não informativos.

[11] A distribuição gamma para precisão e a gamma inversa para variância são utilizadas por duas razões. Primeiro, a precisão (ou variância) precisa assumir apenas valores positivos. Assim, qualquer função de densidade de probabilidade que tenha apenas valores positivos pode ser usada para precisão e variância das priores. Para variáveis contínuas, tais distribuições incluem uma distribuição uniforme que é restrita aos números positivos e a distribuição gamma. Esta é um tanto mais flexível que aquela e permite uma melhor incorporação do conhecimento *a priori*.

Segundo, antes do uso de computação de alta velocidade, a maioria das análises bayesianas era realizada utilizando análises conjugadas. Em uma análise conjugada, a distribuição de probabilidades *a priori* é solicitada de maneira que tenha a mesma forma que a distribuição de probabilidades *a posteriori*. Essa conveniente propriedade matemática permite uma solução de forma fechada (analítica) para a complexa integração envolvida no Teorema de Bayes. Para dados que seguem a distribuição normal, a priore conjugada para o parâmetro que especifica a média é uma distribuição normal, e *a priori* conjugada para o parâmetro que especifica a precisão (= 1/variância) é uma distribuição gama. Para dados que seguem a distribuição de Poisson, a distribuição gamma é a priore conjugada para o parâmetro que define a média (ou taxa) da distribuição de Poisson. Para uma discussão adicional, *ver* Gelman et al. (1995).

A distribuição gamma possui dois parâmetros, escritos como $\Gamma(a,b)$, onde a é tratado como o parâmetro de forma e b é o parâmetro de escala. A função densidade de probabilidade da distribuição gamma é:

(*Continua*)

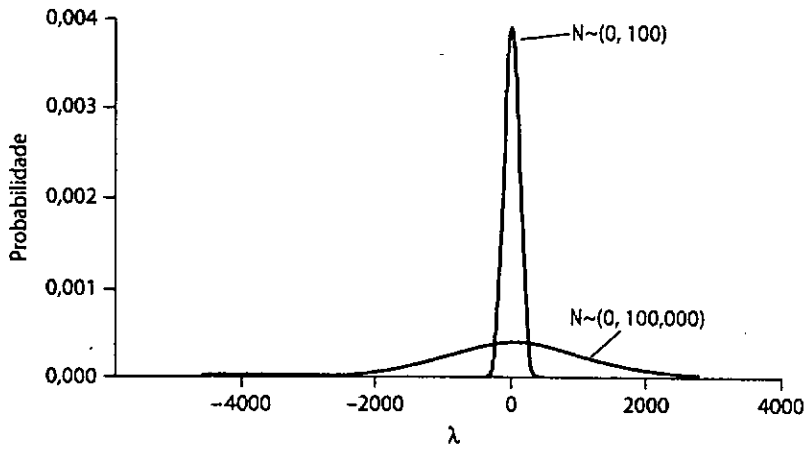


Figura 5.6 Distribuição de probabilidades *a priori* para análise bayesiana. Essa análise requer a especificação da distribuição de probabilidades *a priori* para os parâmetros estatísticos de interesse. Na análise dos dados da Tabela 5.1, o parâmetro é a densidade média de formigueiros, λ. Começamos com uma simples distribuição de probabilidades *a priori* não informativa de que a densidade média de formigueiros é descrita por uma distribuição normal com média 0 e desvio-padrão de 100.000 (curva verde). Como o desvio-padrão é muito grande, a distribuição é praticamente uniforme sobre uma grande amplitude de valores entre −1.500 e +1.500, a probabilidade é essencialmente constante (~0,0002), que é apropriada como priore não informativa. A curva preta representa uma distribuição de probabilidade *a priori* mais precisa. Como o desvio-padrão é muito menor (100), a probabilidade não é mais constante ao longo de uma grande amplitude de valores, mas, em vez disso, ela decresce mais acentuadamente com valores extremos.

[11] (*Continuação*)

$$P(X) = \frac{b^a}{\Gamma(a)} X^{(a-1)} e^{-bX}, \text{ para } X > 0$$

onde $\Gamma(a)$ é a função gamma $\Gamma(n) = (n-1)!$ para inteiros $n > 0$. Generalizando, para os números reais z, a função gamma é definida como:

$$\Gamma(z) = \int_0^1 \left[\ln \frac{1}{t}\right]^{z-1} dt$$

A distribuição gamma tem um valor de expectativa $E(X) = a/b$ e variância $= a/b^2$. Duas distribuições comumente utilizadas por estatísticos são casos especiais da distribuição gamma. A distribuição X^2 com ν graus de liberdade é igual a $\Gamma(\nu/2, 0{,}5)$. A distribuição exponencial com parâmetro β discutida no Capítulo 2 é igual a $\Gamma(1,\beta)$.

Por fim, se a variável aleatória $1/X \sim \Gamma(a,b)$, então X é dito como tendo uma distribuição gamma inversa (GI). Para obter priores não informativas da variância de uma variável aleatória normal, tomamos o limite da distribuição GI, com a e b tendendo a 0. Essa é a razão pela qual usamos $a = b = 0{,}001$ como parâmetro *a priori* para a distribuição gamma que descreve a precisão da estimativa.

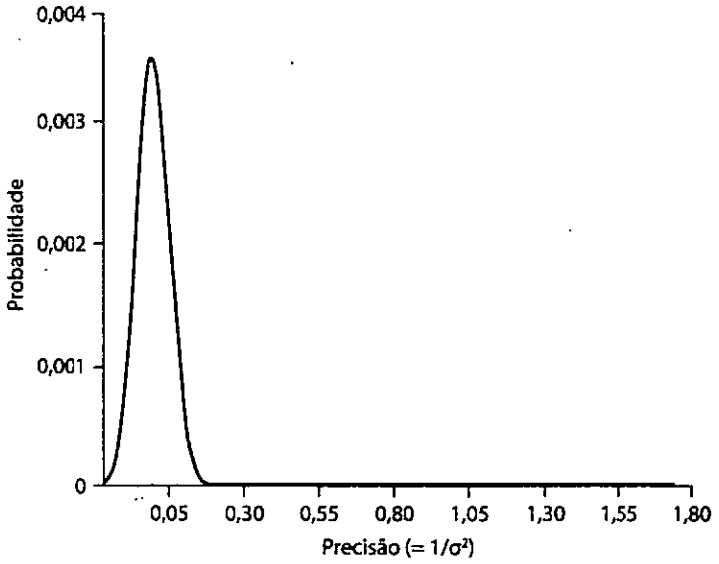


Figura 5.7 Distribuição de probabilidade *a priori* não informativa para a precisão (= 1/variância). A inferência bayesiana requer a especificação não apenas da distribuição de priores para a média da variável (Figura 5.6), mas também uma especificação para a precisão (= 1/variância). A inferência bayesiana em geral especifica uma distribuição gamma inversa para as variâncias. Da mesma forma como com a distribuição das médias, uma priore não informativa é usada para a variância. Neste caso, a variância é distribuída como uma distribuição gamma inversa com parâmetros 1.000 e 1.000: $\sigma^2 \sim GI(1{,}000, 1{,}000)$. Como a variância é muito grande, a precisão é pequena.

Agora temos nossa distribuição de probabilidades *a priori*. As médias das populações desconhecidas de formigas em campos e em florestas, μ_{campo} e $\mu_{floresta}$, são ambas variáveis aleatórias normais com valores de expectativa iguais a λ_{campo} e $\lambda_{floresta}$ e variâncias, σ^2, desconhecidas (mas iguais). Essas expectativas das médias são, por si só, variáveis aleatórias normais com valores esperados de 0 e variância de 100.000. A precisão da população τ é o recíproco da variância da população σ^2, e é uma variável aleatória gamma com parâmetros (0,001, 0,001):

$$\mu_i \sim N(\lambda_i, \sigma^2)$$

$$\lambda_i \sim N(0, 100{,}000)$$

$$\tau = 1/\sigma^2 \sim \Gamma(0.001, 0.001)$$

Essas equações especificam completamente *P*(hipótese) do numerador da Equação 5.2. Se tivéssemos informação *a priori* verdadeira, tal como a densidade de formigueiros em outros campos e florestas, poderíamos usar esses valores para especificar com mais acurácia as médias e variâncias esperadas para os λ.

Passo 4: calculando a verossimilhança

A outra quantidade no numerador do Teorema de Bayes (Equação 5.2) é a verossimilhança, $P(\text{dados}_{obs} \mid \text{hipótese})$. A verossimilhança é uma distribuição proporcional à probabilidade dos dados observados conforme a hipótese.[12] Cada parâmetro λ_i e τ de nossa distribuição de probabilidade *a priori* tem sua própria função de verossimilhança. Em outras palavras, os diferentes valores de λ_i têm funções de verossimilhança que são variáveis aleatórias normais com média igual à observada (aqui, 7 formigueiros por parcela em florestas e 10,75 formigueiros por parcela em campos, *ver* Tabela 5.2). As variâncias são iguais às variâncias das amostras (4,79 em florestas e 2,25 em campos). O parâmetro τ tem uma função de verossimilhança, que é uma variável aleatória gamma inversa. Por último, a razão-F é uma variável aleatória-F com valor de expectativa (ou estimativa de **máxima verossimilhança**)[13] igual a 8,78 (calculada a partir dos dados usando a Equação 5.1).

Passo 5: calculando a distribuição de probabilidades *a posteriori*

Para calcular a distribuição de probabilidades *a posteriori*, $P(\text{H} \mid \text{dados})$, aplicamos a Equação 5.2, multiplicamos a priore pela verossimilhança, e dividimos pela constante de normalização (ou verossimilhança marginal). Embora essa multiplicação seja simples para distribuições com forma de sino, como a normal, os métodos computacionais são usados para estimar iterativamente a distribuição *a posteriori* para qualquer distribuição de priores (Carlin e Louis, 2000).

---

[12] Existe uma diferença-chave entre função de verossimilhança e distribuição de probabilidade. A probabilidade dos dados dada uma hipótese, $P(\text{dados} \mid \text{hipótese})$, é a de qualquer conjunto de dados aleatórios de acordo com uma hipótese específica, em geral a nula estatística. A função densidade de probabilidade associada (*ver* Capítulo 2) conforma com o Primeiro Axioma da Probabilidade – de que a soma das probabilidades = 1. Em contraste, a verossimilhança é fundamentada em apenas um conjunto de dados (a amostra observada) e pode ser calculada para muitas hipóteses e parâmetros diferentes. Apesar de ser uma função que resulta em uma distribuição de valores, a distribuição não é de probabilidades, e a soma de todas as verossimilhanças não necessariamente soma 1.

[13] A máxima verossimilhança é o valor, para o nosso parâmetro, que maximiza a função de verossimilhança. Para obtê-lo, tome a derivada da verossimilhança, configure-o para 0, e a resolva para os valores dos parâmetros. As estimativas de parâmetros frequentistas em geral são iguais às estimativas de máxima verossimilhança dos parâmetros das funções de densidade de probabilidade especificadas. Fisher afirmou, em seu sistema de inferência fiducial, que a estimativa da máxima verossimilhança fornece probabilidades realistas da hipótese alternativa (ou nula). Fisher fundamentou sua afirmação em um axioma estatístico que ele definiu como de acordo com os dados observados Y, a função de verossimilhança $L(\text{H} \mid Y)$ contém todas as informações relevantes acerca da hipótese H. Não discutiremos, aqui, estimativas de máxima verossimilhança, pois (a) as técnicas modernas de computação normalmente as fornecem, em vez de valores assintóticos para as estatísticas frequentistas; e (b) os bayesianos usam a função de verossimilhança inteira em seus cálculos. Para leitura adicional sobre métodos de verossimilhança, *ver* Berger & Wolpert (1984) e Edwards (1992).

Em contraste aos resultados das análises paramétricas e de Monte Carlo, o resultado da análise bayesiana é uma *distribuição* de probabilidades, não apenas um valor de *P*. Dessa forma, neste exemplo, expressamos $P(F \geq 5{,}32 \mid F_{obs})$ como uma *variável aleatória* com média e variância esperadas. Para os dados da Tabela 5.1, calculamos estimativas *a posteriori* para todos os parâmetros: λ_{campo}, $\lambda_{floresta}$, σ^2 (= $1/\tau$) (Tabela 5.7). Como usamos priores não informativas, as estimativas dos parâmetros para as análises bayesiana e paramétrica são similares, embora não idênticas.

A distribuição-F hipotética com valor esperado igual a 5,32 é mostrada na Figura 5.8. Para calcular $P(F \geq 5{,}32 \mid F_{obs})$, simulamos 20.000 razões-F usando um algoritmo de Monte Carlo. A média, ou valor esperado, de todas essas razões-F é 9,77; o número é um pouco maior que a estimativa frequentista (máxima verossimilhança) de 8,78, porque nosso tamanho amostral é muito pequeno ($N = 10$). A dispersão ao redor da média é grande: SD = 7,495; por isso, a precisão de nossa estimativa é relativamente baixa (0,017).

Passo 6: interpretando os resultados

Agora retornamos à questão motivadora. Qual é a probabilidade de obter uma razão-F ≥ 5,32, baseando-nos nos dados sobre a densidade de formigueiros da Tabela 5.1? Em outras palavras, o quanto é provável que as médias da densidade de formigueiros dos dois hábitats realmente sejam diferentes? Podemos responder essa questão de forma direta, perguntando qual a porcentagem de valores na Figura 5.8 é maior ou igual a 5,32. A resposta é 67,3%. Isso não parece muito convincente quanto o valor de $P = 0{,}018$ (1,8%) obtido na análise paramétrica da sessão anterior. De fato, a porcentagem de valores na Figura 5.8 que são ≥ 8,78, o valor observado para o qual encontramos $P = 0{,}018$ na sessão da análise paramétrica, é 46,5. Em outras palavras, a análise bayesiana (Figura 5.8) indica $P = 0{,}67$ de que a densidade de formigueiros nos dois hábitats seja verdadeiramente diferente, dada a estimativa bayesiana de $F = 9{,}77$ [$P(F \geq 5{,}32 \mid F_{obs}) = 0{,}67$]. Em

TABELA 5.7 Estimadores paramétrico e bayesiano para as médias e desvios-padrão dos dados da Tabela 5.1

	Estimadores			
Análise	$\lambda_{floresta}$	λ_{campo}	$\sigma_{floresta}$	σ_{campo}
Paramétrica	7,00	10,75	2,19	1,50
Bayesiana (priore não informativa)	6,97	10,74	0,91	1,13
Bayesiana (priore informativa)	7,00	10,74	1,01	1,02

As estimativas dos desvios-padrão das análises bayesianas são um pouco menores porque a análise bayesiana incorpora informação a partir da distribuição de probabilidades *a priori*. A análise bayesiana pode fornecer resultados diferentes, dependendo da forma da distribuição *a priori* e do tamanho amostral.

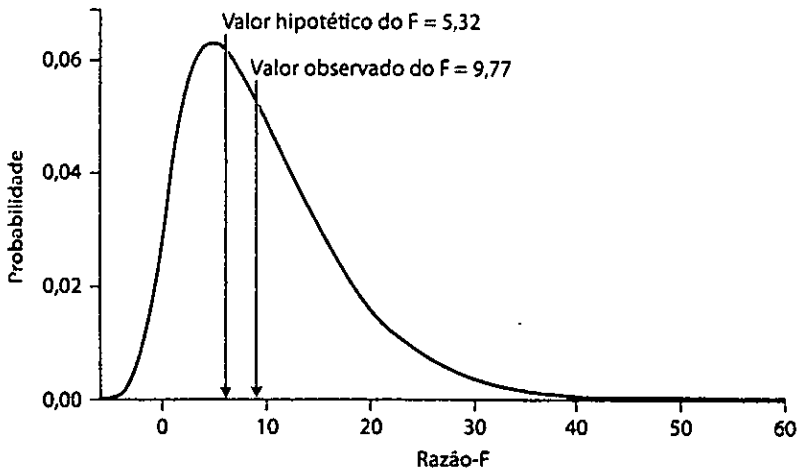


Figura 5.8 Distribuição-F hipotética com um valor esperado de 5,32. Estamos interessados em determinar a probabilidade de que F ≥ 5,32 (o valor crítico para $P < 0,05$ em um teste padrão da razão-F), conforme os dados sobre densidade de formigueiros da Tabela 5.1. Isto é o inverso da hipótese nula tradicional, que questiona: qual é a probabilidade de obter os dados, dada a hipótese nula? Na análise bayesiana, a probabilidade *a posteriori* de F ≥ 5,32, de acordo com os dados da Tabela 5.1, é a proporção da área sob a curva à direita de F = 5,32, que é 0,673. Em outras palavras, *P*(hipótese de que campos e florestas diferem na densidade média de formigueiros | dados observados na Tabela 5.1) = 0,673. O valor mais provável *a posteriori* para a razão-F é 9,77. A proporção da área sob a curva para a direita desse valor é de 0,413. A análise paramétrica diz que os dados observados são improváveis, dada uma distribuição nula especificada pela razão-F {*P*(dados | H_0)}, enquanto a análise bayesiana diz que a probabilidade de observar uma razão-F de 9,77 ou maior não é improvável em razão dos dados [*P*(F ≥ 5,32 | dados) = 0,673].

contraste, a análise paramétrica (Figura 5.5) indica $P = 0,018$ de que a estimativa paramétrica do F = 8,78 (um de uma razão-F maior) poderia ser encontrada dada a hipótese nula de que a densidade de formigueiros nos dois hábitats é a mesma [$P(F_{obs} | H_0) = 0,018$].

Usando a análise bayesiana, uma resposta diferente poderia resultar se tivéssemos usado uma distribuição de priores diferente, em vez de priores não informativas de média 0 e variância muito grande. Por exemplo, se usássemos uma média *a priori* de 15 para as florestas e 7 para campos, uma variância entre grupos de 10, e variância dentro dos grupos de 0,001, então $P(F \geq 5,32 | \text{dados}) = 0,57$. Não obstante, você pode observar que a probabilidade *a posteriori* depende das priores que são utilizadas na análise (Tabela 5.7).

Por fim, podemos estimar um **intervalo de credibilidade** ao redor de nossa estimativa da razão-F observada. Como no método de Monte Carlo, estimamos o intervalo de credibilidade de 95% conforme os 2,5 e 97,5 percentis das razões-F simuladas. Esses valores são 0,28 e 28,39. Dessa forma, podemos

dizer que estamos 95% certos de que o valor da razão-F observado desse experimento está dentro do intervalo [0,28, 28,39]. Note que a dispersão é grande porque a precisão de nossa estimativa da razão-F é baixa, refletida pelo pequeno tamanho amostral em nossa análise. Deveríamos comparar o intervalo de credibilidade com a interpretação do intervalo de confiança apresentado no Capítulo 3.

Premissas da análise bayesiana

Em adição às premissas padrão de todos os métodos estatísticos (observações aleatórias e independentes), a premissa-chave da análise bayesiana é de que os parâmetros a serem estimados são variáveis aleatórias com distribuição conhecida. Em nossa análise, também assumimos pouca informação *a priori* (priores não informativas), portanto, a função de verossimilhança teve mais influência nos cálculos finais da distribuição de probabilidades *a posteriori* do que tiveram as priores. Isso deveria fazer sentido intuitivo. Por outro lado, se tivermos muita informação *a priori*, nossa distribuição de probabilidades *a priori* (p. ex., a curva preta na Figura 5.6) poderia ter baixa variação e a função de verossimilhança não mudaria substancialmente a variância da distribuição de probabilidades *a posteriori*. Se tivéssemos muita informação *a priori*, e confiássemos nela, não aprenderíamos muito com o experimento. Um experimento bem-delineado deveria decrescer a estimativa *a posteriori* da variância em relação à *a priori* da variância.

As contribuições relativas das priores e da verossimilhança para a estimativa *a posteriori* da probabilidade da densidade média de formigueiros em florestas é ilustrada na Figura 5.9. Nesta figura, a priore é achatada ao longo da amplitude dos dados (i. e., ela é uma priore não informativa), a verossimilhança é a distribuição baseada nos dados observados (Tabela 5.1) e a posteriore é o produto das duas. Note que a variância da posteriore é menor que a variância da verossimilhança porque tínhamos alguma informação *a priori*. Contudo, os valores esperados da verossimilhança (7,0) e da posteriore (6,97) estão muito próximos, pois todos os valores da priore eram igualmente prováveis ao longo da amplitude dos dados.

Após fazer esse experimento, temos novas informações sobre cada um dos parâmetros que podem ser utilizados em análises de experimentos subsequentes. Por exemplo, se repetirmos o experimento, poderíamos usar a posteriore na Figura 5.9 como nossa priore para a densidade média de formigueiros em outras florestas. Para isso, poderíamos usar os valores para λ_i e σ_i da Tabela 5.7 como as estimativas de λ_i e σ_i ao estabelecer a nova distribuição de probabilidades *a priori*.

Vantagens e desvantagens da análise bayesiana

A análise bayesiana tem um número de vantagens em relação às abordagens paramétricas e de Monte Carlo conduzidas em uma estrutura frequentista. A análise bayesiana permite a incorporação explícita de informação *a priori*, e os resulta-

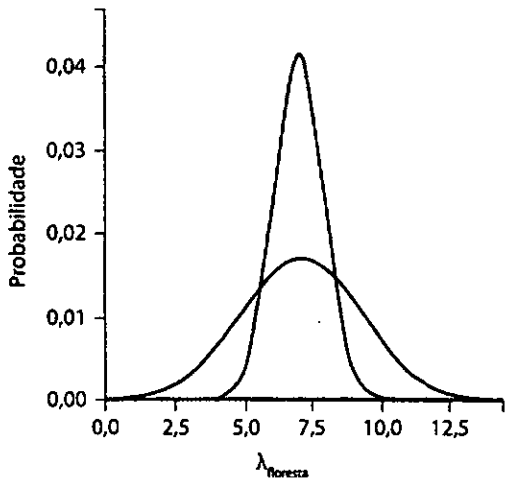


Figura 5.9 Densidade de probabilidades da priore, da verossimilhança e da posteriore para o número médio de formigueiros das parcelas de florestas. Na análise bayesiana dos dados da Tabela 5.1, usamos uma distribuição *a priori* que é não informativa, com média 0 e variância de 100.000. Essa distribuição normal gerou uma amplitude de valores de probabilidade das priores essencialmente uniformes (preta) ao longo da amplitude de valores para $\lambda_{floresta}$, a densidade de formigueiros em florestas. A verossimilhança (verde) representa a probabilidade com base nos dados observados (Tabela 5.1), e a probabilidade *a posteriori* (cinza) é o produto das duas. Note que a distribuição *a posteriori* é mais precisa que a verossimilhança, porque ela leva em consideração a (modesta) informação contida nas priores.

dos de um experimento (*a posteriori*) podem ser usados para informar (como priores) experimentos subsequentes. Os resultados de uma análise bayesiana são interpretados de maneira intuitivamente simples e as inferências obtidas são condicionadas aos dados observados e à informação *a priori*.

As desvantagens da análise bayesiana são seus desafios computacionais (mesmo os programas disponíveis atualmente são difíceis de usar) e o requerimento de condicionar as hipóteses aos dados [i. e., *P*(hipótese | dados)]. A desvantagem mais séria da análise bayesiana é a potencial falta de objetividade, pois diferentes resultados serão obtidos usando-se diferentes priores. Como consequência, diferentes pesquisadores podem obter resultados diferentes a partir do mesmo conjunto de dados, caso eles comecem com diferentes preconceitos ou informações *a priori*. O uso de priores não informativas objetiva solucionar essas críticas, mas aumenta a complexidade computacional.

RESUMO

As três maiores estruturas de análises estatísticas são Monte Carlo, paramétrica e bayesiana. Todas elas assumem que os dados foram amostrados aleatória e independentemente. Na análise de Monte Carlo, os dados são aleatorizados ou rearranjados, de forma que os indivíduos são reatribuídos aos grupos. As estatísticas de teste são calculadas para os conjuntos de dados aleatorizados e o rearranjo é

repetido muitas vezes para gerar uma distribuição de valores simulados. A probabilidade em cauda da estatística de teste é, então, estimada a partir dessa distribuição. As vantagens da análise de Monte Carlo são: não faz nenhuma premissa acerca da distribuição dos dados, torna a hipótese nula clara e explícita, e pode ser relacionada a conjuntos de dados e hipóteses individuais. A desvantagem é que ela não é uma solução geral e muitas vezes requer uma programação de computador para ser implementada.

Na análise paramétrica, os dados são assumidos como tendo sido tirados de uma distribuição subjacente conhecida. Uma estatística-teste observada é comparada à distribuição teórica com base na hipótese nula de variação aleatória. A vantagem da análise paramétrica é que ela fornece uma estrutura unificadora para os testes estatísticos de hipóteses nulas clássicas. A análise paramétrica também é familiar à maioria dos ecólogos e cientistas do ambiente e é amplamente implementada em *softwares* estatísticos. Sua desvantagem é que os testes não especificam a probabilidade da hipótese alternativa, que, com frequência, têm maior interesse que a hipótese nula.

A análise bayesiana considera os parâmetros como sendo variáveis aleatórias se considerados os valores fixos. Ela pode tomar vantagem explícita de informações *a priori*, apesar dos métodos bayesianos modernos apoiarem-se em priores não informativas. Os resultados das análises bayesianas são expressos como distribuição de probabilidades, e suas interpretações conformam a nossa intuição. Contudo, a análise bayesiana requer computação complexa e frequentemente requer que o pesquisador escreva seu próprio programa.

PARTE II

DELINEAMENTO DE EXPERIMENTOS

6

Delineando Estudos de Campo com Sucesso

A análise de dados adequada está relacionada com o delineamento amostral e o esboço experimental adequados. Se existem erros ou problemas sérios no delineamento do estudo ou na coleta de dados, raras vezes é possível repará-los após o erro. Em contraste, se o estudo é devidamente delineado e executado, os dados com frequência podem ser analisados de várias formas, a fim de responder diferentes questões. Neste capítulo discutimos uma vasta gama de assuntos que você precisa considerar ao delinear um estudo em ecologia. Não precisamos superestimar a importância de pensar sobre esses assuntos *antes* de começar a coletar dados.

QUAL É A QUESTÃO CENTRAL DO ESTUDO?

Apesar de parecer jocoso e a resposta óbvia, muitos estudos são iniciados sem uma resposta clara para esta questão. A maioria das respostas irá tomar a forma de questões mais específicas.

Existem diferenças espaciais ou temporais na variável *Y*?

Em geral esta questão é respondida com a coleta de dados e ela representa o ponto de partida de muitos estudos ecológicos. Os métodos estatísticos padrão, como a análise de variância (ANOVA) e a regressão, são bem adequados para responder esta questão. Além disso, o teste e a rejeição convencional de uma hipótese nula simples (*ver* Capítulo 4) produzem uma resposta dicotômica de sim ou não para essa pergunta. É difícil até mesmo discutir os mecanismos sem ter nenhuma noção dos padrões espaciais e temporais em seus dados. Entender as forças que controlam a diversidade biológica, por exemplo, requer, no mínimo, um mapa espacial da riqueza de espécies. O delineamento e a implementação de uma amostragem em ecologia com sucesso requer grande esforço e cuidado, o que também

é necessário para um estudo experimental com sucesso. Em alguns casos, um estudo amostral irá endereçar todos os seus objetivos de pesquisa; em outros, um estudo amostral será o primeiro passo em um projeto de pesquisa. Uma vez que tenha documentado os padrões espaciais e temporais em seus dados, você conduzirá experimentos ou coletas adicionais para acessar os mecanismos responsáveis por aqueles padrões.

Qual é o efeito do fator *X* sobre a variável *Y*?

Esta é a questão diretamente respondida por um experimento manipulativo. Em um experimento de campo ou de laboratório, o pesquisador estabelece ativamente diferentes níveis do fator *X* e mede a resposta da variável *Y*. Se o delineamento experimental e as análises estatísticas são apropriados, o valor de *P* resultante pode ser utilizado para testar a hipótese nula de nenhum efeito do fator *X*. Resultados estatisticamente significativos sugerem que o fator *X* influencia a variável *Y*, e que o "sinal" do fator *X* é forte o suficiente para ser detectado acima do "ruído" causado por outras fontes de variação natural.[1] Certos experimentos naturais podem ser analisados da mesma forma, tirando vantagem da variação natural que existe no fator *X*. Contudo, as inferências resultantes geralmente são fracas, pois há menos controle sobre variáveis confundidas. Discutiremos experimentos naturais com mais detalhe adiante, neste capítulo.

As medidas da variável *Y* são consistentes com as predições da hipótese H?

Esta pergunta representa o clássico confronto entre teoria e dados (Hilborn & Mangel, 1997). No Capítulo 4, discutimos duas estratégias usadas neste confronto: a abordagem indutiva, na qual uma única hipótese é recursivamente modificada para se adaptar aos dados que se acumulam; e a abordagem hipotético-dedutiva, na qual as hipóteses são falseadas e descartadas caso não predigam os dados. Dados de estudos experimentais ou observacionais podem ser utilizados para questionar se as observações são consistentes com as predições de uma hipótese mecanística. Infelizmente, os ecólogos nem sempre postulam essa questão tão claramente. Suas duas limitações são: (1) muitas hipóteses ecológicas não geram predições simples e falseáveis e (2) mesmo quando uma hipótese gera predições, raras vezes são únicas. Dessa forma, pode não ser possível testar definitivamente a Hipótese H usando apenas dados coletados sobre a variável *Y*.

[1] Embora os experimentos manipulativos permitam inferências fortes, eles podem não revelar mecanismos explícitos. Muitos experimentos na ecologia são simplesmente uma espécie de "caixa preta", que medem a *resposta* da variável *Y* a mudanças no fator *X*, mas não elucidam mecanismos em níveis inferiores que causam aquela resposta. Tal entendimento mecanístico pode exigir observações ou experimentos adicionais, objetivando uma questão mais específica acerca do processo.

Usando as medidas da variável *Y*, qual é a melhor estimativa do parâmetro θ no modelo Z?

Modelos estatísticos e matemáticos são ferramentas poderosas em ecologia e em ciências ambientais. Eles permitem a previsão de como as populações e as comunidades vão mudar ao longo do tempo ou responder a condições ambientais alteradas (p. ex., Sjögren-Gulve & Ebenhard, 2000). Os modelos também podem nos ajudar a entender como diferentes mecanismos ecológicos interagem de maneira simultânea para controlar a estrutura de comunidades e populações (Caswell, 1988). A estimativa de parâmetros é necessária para construir modelos preditivos e é uma característica especialmente importante das análises bayesianas (*ver* Capítulo 5). Raras vezes há uma correspondência simples, de um pra um, entre o valor da variável *Y*, medida em campo, e o valor do parâmetro θ em nosso modelo. Em vez disso, esses parâmetros devem ser extraídos e estimados indiretamente a partir de nossos dados. Infelizmente, alguns dos mais comuns e tradicionais delineamentos utilizados em experimentos em ecologia e amostragens de campo, como a análise de variância (*ver* Capítulo 10), não são muito úteis para estimar parâmetros de modelos. No Capítulo 7, discutiremos alguns delineamentos alternativos mais úteis na estimativa de parâmetros.

EXPERIMENTOS MANIPULATIVOS

Em um **experimento manipulativo**, o pesquisador primeiro altera os níveis da variável preditora (ou o fator) e então mede como uma ou mais variáveis de interesse respondem a essas alterações. Tais resultados são então utilizados para testar hipóteses de causa e efeito. Por exemplo, se estamos interessados em testar a hipótese de que a predação por lagartos controla a densidade de aranhas em pequenas ilhas do Caribe, podemos alterar a densidade de lagartos em uma série de cercados e medir a densidade de aranhas resultante (p. ex., Spiller & Schoener, 1998). Poderíamos, então, representar graficamente estes dados, de maneira que o eixo-*x* (= variável independente) é a densidade de lagartos e o eixo-*y* (= variável dependente), a densidade de aranhas (Figura 6.1A,B).

Nossa hipótese nula é de que não há relação entre essas duas variáveis (Figura 6.1A). Isto é, a densidade de aranhas pode ser alta ou baixa em um cercado particular, mas não é relacionada com a de lagartos, que foi colocada no cercado. Além disso, podemos observar uma relação negativa entre a densidade de aranhas e de lagartos: cercados com as maiores densidades de lagartos têm menos aranhas e vice-versa (Figura 6.1B). Logo, esse padrão pode ser sujeito a uma análise estatística, como a regressão (Capítulo 9), para determinar, ou não, se a evidência foi suficiente para rejeitar a hipótese nula de que não há relação entre as densidades de lagartos e de aranhas. A partir desses dados também podemos estimar os parâmetros do modelo de regressão que quantificam a força da relação.

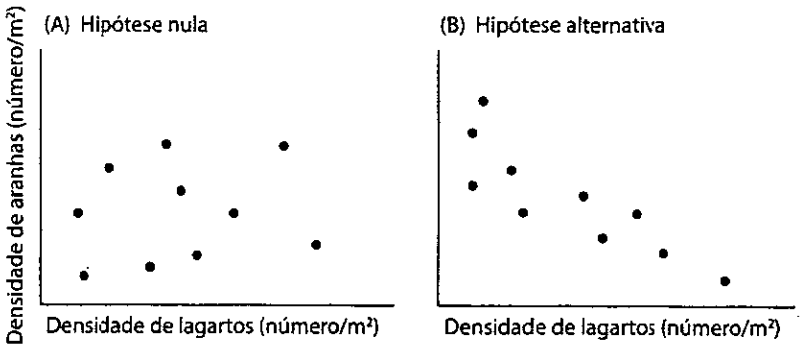


Figura 6.1 Relação entre a densidade de lagartos e a de aranhas em um experimento de campo natural e manipulativo. Cada ponto representa uma parcela ou um cercado em que ambas, densidade de aranhas e de lagartos, foram medidas. (A) A hipótese nula é de que a densidade de lagartos não tem efeito sobre a de aranhas. (B) A hipótese alternativa é de que a predação por lagartos controla a densidade de aranhas, levando a uma relação negativa entre essas duas variáveis.

Apesar de os experimentos de campo serem populares e poderosos, eles possuem várias limitações importantes. Primeiro, é desafiador conduzir experimentos em grandes escalas espaciais; mais de 80% dos experimentos de campo foram conduzidos em parcelas com menos de 1 m^2 (Kareiva & Anderson, 1988; Wiens, 1989). Quando experimentos são conduzidos em grandes escalas espaciais, a replicação é inevitavelmente sacrificada (Carpenter, 1989). Mesmo quando replicado adequadamente, experimentos conduzidos em pequenas escalas espaciais podem não produzir resultados que sejam representativos dos padrões e processos que ocorrem em escalas espaciais maiores (Englund & Cooper, 2003).

Segundo, experimentos de campo frequentemente são restritos a organismos de corpo pequeno e de vida curta, mais fáceis de manipular. Apesar da tendência que temos em generalizar os resultados de nossos experimentos para outros sistemas, é improvável que a interação entre lagartos e aranhas nos diga muito sobre a interação entre leões e gnus. Terceiro, é difícil mudar uma, e apenas uma, variável por vez em experimentos manipulativos. Por exemplo, gaiolas podem excluir outros tipos de predadores e presas, além de aumentar o sombreamento. Se compararmos cuidadosamente a densidade de aranhas em gaiolas fechadas *versus* gaiolas abertas "controle", os efeitos da predação por lagartos são **confundidos** com outras diferenças físicas entre os tratamentos. Mais adiante, neste capítulo, discutiremos soluções para variáveis confundidas.

Por último, muitos delineamentos experimentais padrão são simplesmente desajeitados para experimentos de campo realísticos. Por exemplo, suponha que estamos interessados em investigar interações competitivas em um grupo de oito espécies de aranhas. Cada tratamento em tal experimento poderia consistir em uma combinação única de espécies. Embora o número de espécies em cada tratamento varie de 1 a 8, o de combinações únicas é $2^8 - 1 = 255$. Se

quisermos estabelecer 10 réplicas de cada tratamento (*ver* "A Regra dos 10", discutida posteriormente neste capítulo), necessitaríamos de 2.550 parcelas. Isto pode não ser possível por causa de restrição de espaço, tempo ou mão-de-obra. Por causa de todas essas limitações em potencial, muitas questões importantes em ecologia de comunidades não podem ser endereçadas com experimentos de campo.

EXPERIMENTOS NATURAIS

Um **experimento natural** (Cody, 1974) realmente não é um experimento completo. Pelo contrário, é um estudo observacional no qual tiramos vantagem da variação natural presente na variável de interesse. Por exemplo, em vez de diretamente manipular a densidade de lagartos (uma empreitada cara, difícil e demorada), poderíamos fazer o censo em um conjunto de parcelas (ou ilhas) que variam naturalmente na densidade de lagartos (Schoener, 1991). Idealmente, essas parcelas devem variar *apenas* na densidade de lagartos e serem idênticas em todos os outros aspectos. Poderíamos, então, analisar a relação entre a densidade de aranhas e a de lagartos, como ilustrado na Figura 6.1.

Experimentos naturais e experimentos manipulativos em geral resultam nos mesmos tipos de dados e são analisados com os mesmos tipos de estatísticas. Porém, com frequência existem diferenças importantes na interpretação de experimentos naturais e manipulativos. Em um experimento manipulativo, se tivermos estabelecido controles válidos e mantido as mesmas condições ambientais entre as réplicas, qualquer diferença consistente na variável resposta (p. ex., densidade de aranhas) pode ser atribuída, com confiança, às diferenças no fator manipulado (p. ex., densidade de lagartos).

Não temos a mesma confiança ao interpretar resultados de experimentos naturais. Nesse caso, não sabemos a direção da causa e efeito, e não controlamos outras variáveis que certamente variam entre as réplicas. Para o exemplo dos lagartos e aranhas, há pelo menos quatro hipóteses capazes de explicar uma associação negativa entre a densidade de lagartos e aranhas:

1. Lagartos podem controlar a densidade de aranhas. Eis a hipótese alternativa de interesse do experimento de campo original.
2. Aranhas podem, direta ou indiretamente, controlar a densidade de lagartos. Suponha, por exemplo, que grandes aranhas caçadoras consomem pequenos lagartos, ou que aranhas também sejam predadas por pássaros que igualmente se alimentam de lagartos. Em ambos os casos, o aumento na densidade de aranhas pode decrescer a de lagartos; mesmo que estes comam aranhas.
3. Ambas, densidades de aranhas e de lagartos, são controladas por um fator ambiental não medido. Por exemplo, suponha que a densidade de aranhas é maior em parcelas úmidas e a densidade de lagartos, em parcelas secas. Mesmo que os lagartos tenham pouco efeito sobre as aranhas, o padrão mostrado

na Figura 6.1B irá aparecer: parcelas úmidas terão muitas aranhas e poucos lagartos; e parcelas secas, o contrário.

4. Fatores ambientais podem controlar a força da interação entre aranhas e lagartos. Por exemplo, eles podem ser eficientes predadores de aranhas em parcelas secas, mas ineficientes em parcelas úmidas. Em tais casos, a densidade de aranhas dependerá tanto da de lagartos quanto do nível de umidade da parcela (Spiller & Schoener, 1995).

Esses quatro cenários são apenas os mais simples que podem levar a uma relação negativa entre a densidade de lagartos e a de aranhas (Figura 6.2). Se adicionarmos setas de duas pontas ao diagrama (aranhas e lagartos afetam reciprocamente a densidade um do outro), há uma gama ainda maior de hipóteses que podem explicar a relação observada entre a densidade de lagartos e a de aranhas (Figura 6.1).

Entretanto, tudo isso não quer dizer que não há esperança para experimentos naturais: em muitos casos podemos coletar dados adicionais para distinguir entre essas hipóteses. Por exemplo, se suspeitarmos que as variáveis ambientais, como a umidade, são importantes, podemos tanto restringir a amostragem a um conjunto de parcelas com níveis de umidade comparáveis, ou (melhor ainda) medir a densidade de lagartos, a de aranhas e os níveis de umidade em uma série de parcelas ao longo de um gradiente de umidade. Variáveis confundidas e mecanismos alternativos também podem ser problemáticos em experimentos de campo. Contudo, seus impactos serão minimizados se o investigador conduzir experimentos em escalas espaciais e temporais adequadas, estabelecer controles apropriados,

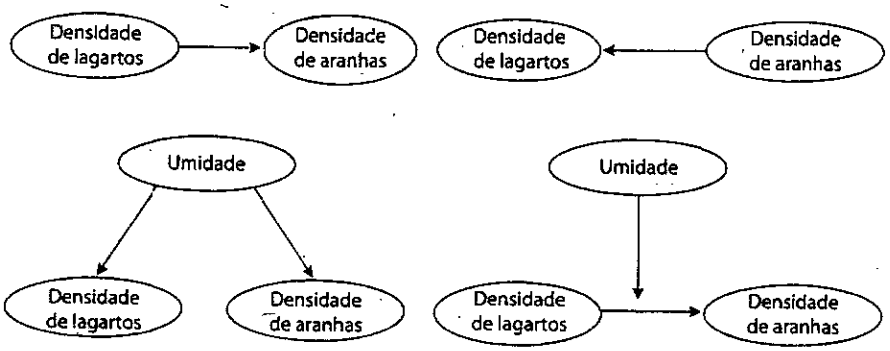


Figura 6.2 Hipóteses mecanísticas para explicar as correlações entre a densidade de lagartos e a de aranhas (Figura 6.1). A relação de causa e efeito pode ser a partir do predador para a presa (esquerda superior) ou da presa para o predador (direita superior). Modelos mais complicados incluem os efeitos de outras variáveis bióticas ou abióticas. Por exemplo, pode não haver interação entre lagartos e aranhas, mas a densidade de ambos pode ser controlada por uma terceira variável, como a umidade (esquerda inferior). Alternativamente, a umidade pode ter um efeito indireto, alterando a interação de lagartos e aranhas (direita inferior).

replicar adequadamente e usar a randomização para estabelecer a posição das réplicas e para atribuir os tratamentos.

Definitivamente, experimentos manipulativos permitem maior confiança em nossas inferências sobre causa e efeito, mas são limitados a escalas espaciais relativamente pequenas e curtas estruturas de tempo. Experimentos naturais podem ser conduzidos em, virtualmente, qualquer escala espacial (de pequenas parcelas a continentes inteiros) e ao longo de qualquer intervalo de tempo (de medidas de campo semanais, a censos anuais até estratos fósseis). Contudo, é mais desafiador distinguir entre relações de causa e efeito em experimentos naturais.[2]

EXPERIMENTOS FOTOGRÁFICOS *VERSUS* EXPERIMENTOS DE TRAJETÓRIA

As duas variantes de experimentos naturais são: os experimentos fotográficos e os de trajetória (Diamond, 1986). Os primeiros são replicados no espaço e os de trajetória, no tempo. Para os dados da Figura 6.1, suponha que fizemos o censo em 10 parcelas diferentes em um único dia. Este é um experimento fotográfico no qual a replicação é espacial; cada observação representa uma parcela diferente em que o censo foi feito ao mesmo tempo. Por outro lado, suponha que visitamos uma mesma parcela em 10 anos diferentes. Trata-se de um experimento de trajetória cuja replicação é temporal; cada observação representa um ano diferente no estudo.

As vantagens de um experimento fotográfico é que ele é rápido, e as réplicas espaciais são estatisticamente mais independentes umas das outras do que as réplicas temporais de um experimento de trajetória. A maioria dos conjuntos de dados ecológicos é de experimentos fotográficos, refletindo a estrutura de 3 a 5 anos oferecida pela maioria dos financiamentos de pesquisa e teses de doutorado.[3] De fato, muitos estudos de mudança temporal são, na realidade, experimentos fotográficos, pois a variação no espaço é tratada como variável substituta para

[2] Em alguns casos, a distinção entre experimentos manipulativos e experimentos naturais de campo não é trivial. A atividade humana tem gerado muitos experimentos não intencionais em larga escala, incluindo eutrofização, alteração de hábitats, mudança climática global e introdução ou remoção de espécies. Ecólogos criativos podem tirar vantagem dessas alterações para delinear estudos em que a confiança nas conclusões é muito alta. Por exemplo, Knapp et al. (2001) estudaram os impactos da introdução de trutas em lagos na Serra Nevada, comparando comunidades de invertebrados em lagos naturalmente sem peixes, lagos com peixes e lagos que já tiveram peixes. Muitas comparações desse tipo são possíveis para documentar as consequências da atividade humana. Contudo, conforme os impactos humanos se tornam mais disseminados e perceptíveis, pode ser cada vez mais difícil encontrar locais que possam ser considerados controles sem manipulação.

[3] Uma exceção notável a experimentos ecológicos de curta duração é o conjunto de estudos coordenados desenvolvidos nas áreas de Pesquisa Ecológica de Longa Duração (PELD). A National Science Foundation – NSF (Estados Unidos) financiou o estabelecimento desses locais ao longo dos anos 1980 e 90, especificamente para visar às necessidades de estudos de pesquisas ecológicas que alcancem de décadas a séculos. Ver http://www.lternet.edu/.

a variação no tempo. Por exemplo, a mudança de sucessão em comunidades de plantas pode ser estudada amostrando uma cronossequência – um conjunto de observações, locais ou hábitats ao longo de um gradiente espacial que difere no tempo de origem (p. ex., Law et al., 2003).

A vantagem de um experimento de trajetória é que ele revela como os sistemas ecológicos mudam ao longo do tempo. Muitos modelos ecológicos e ambientais descrevem precisamente o tipo de mudança, e esse tipo de experimento permite comparações mais fortes entre as predições do modelo e os dados de campo. Além disso, muitos modelos para conservação ou de previsão ambiental são elaborados para se prever condições futuras, e dados para esses modelos são derivados com mais confiança de experimentos de trajetória. Muitos dos conjuntos de dados mais valiosos da ecologia são de longas séries temporais, nos quais populações e comunidades em um local são amostradas ano após ano com métodos consistentes e padronizados. Contudo, esses experimentos que são restritos a um único local não são replicados no espaço. Não sabemos se as trajetórias temporais descritas para aquele local são típicas para o que podemos encontrar em outros lugares. Cada trajetória é essencialmente de tamanho amostral *um* em um dado local.[4]

O problema da dependência temporal

Um problema mais difícil com os experimentos de trajetória é a potencial não independência dos dados coletados em uma sequência temporal. Por exemplo, suponha que você meça o diâmetro de árvores todo mês por um ano em uma parcela de sequóias. Sequoias crescem muito devagar, logo as medidas de um mês para outro serão idênticas. Muitos engenheiros florestais diriam que você não tem 12 amostras independentes, você tem apenas uma (o diâmetro médio para aquele ano). Por outro lado, medidas mensais de uma comunidade planctônica de água doce e de rápido desenvolvimento podem ser vistas como estatisticamente independentes uma da outra. Naturalmente, quanto mais afastadas no tempo uma amostra for de outra, mais elas funcionarão como réplicas independentes.

Mesmo se o intervalo correto entre os censos for empregado, ainda há um súbito problema em como a mudança temporal deve ser modelada. Por exemplo, suponha que você está tentando modelar a mudança no tamanho populacional de uma planta anual de deserto, para o qual você tem acesso a um bom estudo de

[4] Tanto os delineamentos instantâneos quanto os de trajetória são vistos em experimentos manipulativos. Em particular, alguns delineamentos incluem uma série de medidas realizadas antes e depois de uma manipulação. As medidas "prévias" servem como um tipo de "controle", que pode ser comparado às realizadas após a manipulação ou intervenção. Esse tipo de delineamento ADCI (Antes-Depois; Controle-Impacto) é especialmente importante em análises de impacto ambiental e em estudos onde a replicação espacial é limitada. Para mais informações sobre ADCI, *ver* a seção "Estudos em Larga Escala e Impactos Ambientais" mais adiante, neste Capítulo, e no Capítulo 7.

trajetória, com 100 anos de censos anuais consecutivos. Você poderia ajustar um modelo de regressão linear padrão (*ver* Capítulo 9) para a série temporal.

$$N_t = \beta_0 + \beta_1 t + \varepsilon \qquad (6.1)$$

Nesta equação, o tamanho populacional (N_t) é uma função linear da quantidade de tempo (t) que passou. Os coeficientes β_0 e β_1 são o intercepto e a inclinação dessa linha reta. Se β_1 é menor que 0,0, a população está encolhendo com o tempo; se $\beta_1 > 0$, N está aumentando. Aqui, ε é o termo do erro com distribuição normal chamado de **ruído branco**[5], que incorpora tanto o erro nas medidas e a variação aleatória no tamanho da população. O Capítulo 9 explicará esse modelo em muito mais detalhes, mas o apresentamos agora como uma forma simples de pensar acerca de como o tamanho populacional pode mudar de forma linear com o passar do tempo.

Contudo, esse modelo não leva em consideração que o tamanho populacional muda através de nascimentos e mortes, afetando o tamanho populacional *atual*. Um **modelo de série temporal** pode descrever o crescimento populacional como

$$N_{t+1} = \beta_0 + \beta_1 N_t + \varepsilon \qquad (6.2)$$

Neste modelo, o tamanho da população no próximo passo de tempo (N_{t+1}) não depende simplesmente da quantidade de tempo t que passou, mas sim do tamanho populacional no último passo de tempo (N_t). Nesse modelo, a constante β_1 é um termo multiplicador que determina se a população está aumentando exponencialmente ($\beta_1 > 1{,}0$) ou diminuindo ($\beta_1 < 1{,}0$). Como antes, ε é um termo de erro de ruído branco.

O modelo linear (Equação 6.1) descreve um simples aumento *aditivo* do N com o tempo, enquanto a série temporal, ou modelo **autoregressivo** (Equação 6.2), descreve um aumento *exponencial*, pois o fator β_1 é um multiplicador que, em média, dá uma porcentagem constante de aumento no tamanho populacional a cada passo no tempo. A diferença mais importante entre os dois modelos, con-

[5] Ruído branco é um tipo de distribuição de erro na qual os erros são independentes e não correlacionados uns com os outros. Ele é chamado de ruído branco como analogia à luz branca, que é uma mistura igual de comprimentos de onda longos e curtos. Em contraste, a distribuição de ruído vermelho é dominada por perturbações de baixa frequência, assim como a luz vermelha é dominada por ondas de luz de baixa frequência. A maioria das séries temporais, se avaliado o tamanho de populacional, exibe um ruído de espectro avermelhado (Pimm & Redfearn, 1988), de forma que as variâncias no tamanho populacional aumentam quando elas são analisadas em escalas temporais maiores. Modelos de regressão paramétrica requerem termos de erro com distribuição normal, de forma que as distribuições de ruído branco formam as bases para a maioria dos modelos ecológicos estocásticos. Contudo, um espectro inteiro de distribuições de ruído coloridos (ruído 1/f) podem fornecer um ajuste melhor a muitos conjuntos de dados de ecologia e evolução (Halley, 1996).

tudo, é que as diferenças entre o tamanho populacional observado e o predito (i. e., os **desvios**) no modelo de série temporal são correlacionados uns com os outros. Como consequência, há uma tendência em haver períodos de aumento consecutivos seguidos por períodos de diminuições consecutivas. Isto ocorre porque a trajetória de crescimento tem uma "memória" – cada observação consecutiva (N_{t+1}) depende intimamente da que veio antes (o termo N_t na Equação 6.2). Em contraste, o modelo linear não tem memória, e os aumentos são função apenas do tempo (e do ε) e não de N_t. Portanto, os desvios positivos e negativos seguem um ao outro em uma tendência puramente aleatória (Figura 6.3). Desvios correlacionados, que são típicos em dados coletados em estudos de trajetória, violam as premissas da maioria das análises estatísticas mais convencionais.[6] Novos métodos analíticos e computacionalmente intensivos têm sido desenvolvidos para analisar tanto dados amostrais quanto dados experimentais coletados através do tempo (Ives et al., 2003; Turchin, 2003).

Isso não quer dizer que não podemos incorporar dados de séries temporais em análises estatísticas convencionais. Nos Capítulos 7 e 10, discutiremos formas adicionais para analisar dados de séries temporais. Esses métodos demandam que você tenha uma atenção cuidadosa tanto no delineamento amostral quanto no tratamento dos dados após a coleta. Sobre isso, dados de séries temporais ou de trajetória são iguais a qualquer outro tipo de dado.

EXPERIMENTOS DE PRESSÃO *VERSUS* EXPERIMENTOS DE PULSOS

Em experimentos manipulativos, também distinguimos entre **experimentos de pressão** e **experimentos de pulsos** (Bender et al., 1984). No primeiro caso, as condições alteradas no tratamento são mantidas ao longo do tempo e são reaplicadas conforme a necessidade, para assegurar que a força da manipulação permaneça constante. Assim, o fertilizante talvez deva ser reaplicado às plantas, e animais que morreram ou desapareceram de uma parcela podem ter de ser repostos. Ao contrário, em um experimento de pulsos, os tratamentos experimentais são aplicados apenas uma vez, no começo do estudo. O tratamento não é reaplicado e à réplica é permitida que se recupere da manipulação (Figura 6.4A).

Os experimentos de pressão e de pulso medem duas respostas diferentes ao tratamento. O de pressão (Figura 6.4B) mede a resistência do sistema ao trata-

[6] Na verdade, a autocorrelação espacial gera os mesmos problemas (Legendre & Legendre, 1998; Lichstein et al., 2003). Contudo, ferramentas para a análise de autocorrelação espacial têm sido desenvolvidas mais ou menos independentemente das análises de séries temporais, talvez porque percebemos o tempo como uma variável estritamente unidimensional e o espaço como uma variável bi ou tridimensional.

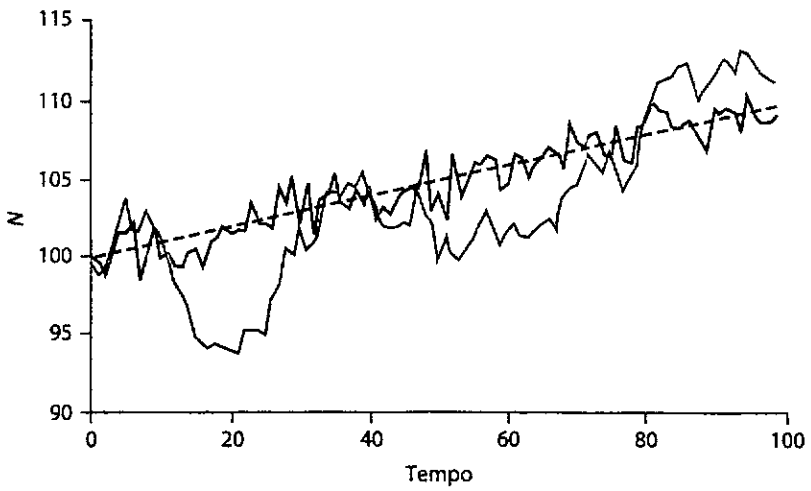


Figura 6.3 Exemplos de séries temporais determinísticas e estocásticas, com e sem autocorrelação. Cada população começa com 100 indivíduos. Um modelo linear sem erro (linha tracejada) ilustra uma tendência de crescimento constante na população. Um modelo linear com termo de erro de ruído branco estocástico (linha preta) adiciona variabilidade temporal não correlacionada. Por último, um modelo autocorrelacionado (linha verde) descreve o tamanho da população no próximo passo de tempo (t+1), como função do tamanho populacional no tempo atual (t), acrescido de ruído aleatório. Apesar de o termo de erro nesse modelo ainda ser uma simples variável aleatória, a série temporal resultante apresenta autocorrelação – existem períodos de crescimento populacional seguidos de declínio da população. Para o modelo linear e para o modelo de ruído branco estocástico, a equação é $N_t = a + bt + \varepsilon$, onde $a = 100$ e $b = 0{,}10$. Para o modelo autocorrelacionado, $N_{t+1} = a - bN_t + \varepsilon$, onde a = 0,0 e b = 1,0015. Para ambos os modelos com erro, ε é uma variável aleatória normal: $\varepsilon \sim N(0,1)$.

mento experimental: a extensão à qual ele resiste mudar ao ambiente constante criado pelo experimento de pressão. Um sistema com baixa resistência irá exibir uma grande resposta a um experimento de pressão; enquanto isso, um sistema com alta resistência irá exibir pouca diferença entre os tratamentos de controle e manipulados.

Experimentos de pulso medem a resiliência do sistema ao tratamento experimental: a extensão à qual o sistema se recupera de uma única perturbação. Um sistema com alta resiliência irá apresentar um rápido retorno às condições de controle, enquanto um com baixa resiliência levará um longo tempo para se recuperar; parcelas manipuladas e de controle irão continuar diferindo por um longo tempo após uma única aplicação do tratamento.

A distinção entre experimentos de pressão e de pulso não é o número de aplicações dos tratamentos utilizados, mas se as condições alteradas são mantidas ao longo do tempo no tratamento. Se as condições ambientais permanecem constantes após uma única perturbação durante o experimento, o delineamento é efetivamen-

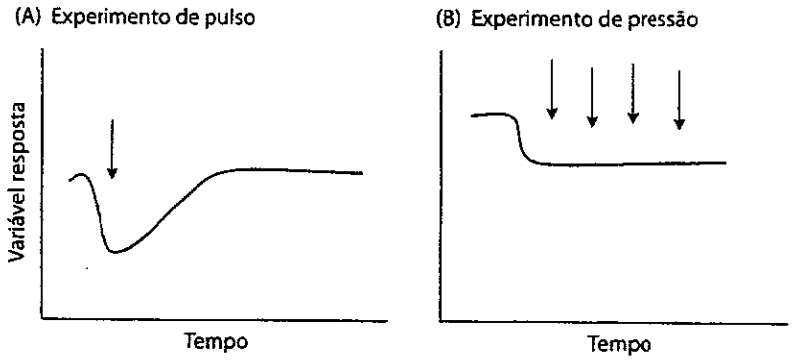


Figura 6.4 Experimentos ecológicos de pulso e pressão. As setas indicam a aplicação do tratamento, e a linha indica a trajetória temporal da variável resposta. O experimento de pulso (A) mede a resposta a uma única aplicação do tratamento (resiliência), enquanto o experimento de pressão (B) mede a resposta sob condições constantes (resistência).

te um experimento de pressão. Outra distinção entre experimentos de pressão e de pulso é que o primeiro mede a resposta do sistema sob condições de equilíbrio, enquanto o segundo registra respostas transitórias em um ambiente em modificação.

REPLICAÇÃO

Quantas réplicas?

Eis uma das questões mais comuns que ecólogos e cientistas da área ambiental perguntam aos estatísticos. A resposta correta é que depende da variância nos dados e do **tamanho do efeito** – a diferença que você deseja detectar entre as médias dos grupos em comparação. Infelizmente, essas duas quantidades são difíceis de estimar, embora você sempre deva considerar que o tamanho do efeito seria razoável observar.

Para estimar variâncias, muitos estatísticos irão recomendar que você conduza um estudo-piloto. Infelizmente, estudos-piloto geralmente são inviáveis – e você raras vezes terá a liberdade de montar e executar um estudo caro e demorado mais de uma vez. Os períodos de campo e as propostas de financiamento são muito curtos para esse tipo de luxúria. Contudo, você deve ser apto a estimar uma amplitude razoável para as variâncias e para o tamanho dos efeitos a partir de estudos publicados previamente e a partir da discussão com colegas. Você pode usar esses valores para determinar o poder estatístico (*ver* Capítulo 4) que irá resultar de diferentes combinações de réplicas, variâncias e tamanhos do efeito (*ver* Figura 4.5, para um exemplo). No mínimo, contudo, você precisa primeiro responder a seguinte questão:

Quantas réplicas são possíveis no total?

Leva tempo, trabalho físico e dinheiro para coletar tanto dados experimentais quanto expedições, e você precisa determinar precisamente o total de amostras

que pode custear. Se você está conduzindo uma cultura ou análise amostral cara, o custo em reais pode ser o fator limitante. Contudo, em muitos estudos, tempo e trabalho físico são mais limitantes que o dinheiro. Isto é especialmente verdade para expedições geográficas conduzidas em grandes escalas espaciais, para as quais você (e sua equipe de campo, se tiver sorte o bastante para ter uma) pode gastar tanto tempo viajando para os locais de estudo quanto coletando os dados de campo. Idealmente, todas as réplicas devem ser medidas de maneira simultânea, proporcionando um perfeito experimento fotográfico. Quanto mais tempo levar para coletar todos os dados, mais as condições terão mudado desde a primeira até última amostra. Para estudos experimentais, se os dados não forem coletados todos de uma vez, então a quantidade de tempo passado desde a aplicação do tratamento não será mais idêntica para todas as réplicas.

Obviamente, quanto maior a escala espacial do estudo, mais difícil é coletar todos os dados dentro de uma estrutura de tempo razoável. Todavia, o saldo pode ser maior, porque o escopo da inferência está atado à escala espacial das análises: conclusões baseadas em amostras tiradas de apenas um local podem não ser válidas para outros locais. Contudo, não há razão para desenvolver um delineamento amostral surreal. Mapeie cuidadosamente seu projeto do começo ao fim para garantir que ele será factível.[7] Uma vez sabendo o número total de réplicas ou observações que você pode coletar, poderá começar a delinear seu experimento, aplicando a Regra dos 10.

A Regra dos 10

A **Regra dos 10** diz que você deve coletar pelo menos 10 réplicas para cada categoria ou nível de tratamento. Por exemplo, suponha que tenha determinado que possa coletar um total de 50 observações em um experimento, examinando taxas fotossintéticas entre diferentes espécies de plantas. Um bom delineamento para uma ANOVA de um fator (*one-way*) poderia ser comparar as taxas fotossintéticas entre não mais que 5 espécies. Para cada uma, você poderá escolher aleatoriamente 10 plantas e tirar uma medida de cada.

A Regra dos 10 não se baseia em nenhum princípio teórico de delineamento experimental ou de análise estatística, mas é uma reflexão de nossa experiência de campo, obtida através de muito esforço, com delineamentos que foram bem-su-

[7] Pode ser bastante informativo usar um cronômetro para medir atenciosamente quanto tempo leva para completar as medidas em uma única réplica de seu estudo. Como o "pai especialista em eficiência", em *Cheaper by the dozen** (Gilbreth & Carey, 1949), damos muita importância a tais números. Com esses dados, podemos estimar com acurácia quantas réplicas poderemos fazer em uma hora, e quanto tempo de campo no total precisaremos para completar o censo. O mesmo princípio aplica-se ao processamento da amostra, medidas que fazemos no laboratório, dados que passamos para o computador e o armazenamento de longo prazo e curadoria dos dados (*ver* Capítulo 8). Todas essas atividades consomem um tempo que deve ser considerado ao planejar um estudo ecológico.

* N. de T. Este livro foi adaptado para o cinema em 1950 e lançado no Brasil com nome de "Papai batuta", sendo regravado e lançado no Brasil em 2003 com o nome de "Doze é demais".

cedidos e aqueles que não foram. Certamente, é possível analisar conjuntos de dados com menos de 10 observações por tratamento, e frequentemente quebramos a regra. Delineamentos balanceados com muitas combinações de tratamentos, mas com apenas quatro ou cinco réplicas, podem ser bastante poderosos. E certos delineamentos de uma via com poucos níveis de tratamentos podem requerer mais de 10 réplicas por tratamento, se as variâncias forem grandes.

Todavia, a Regra dos 10 é um ponto de partida sólido. Mesmo se você estabelecer o delineamento com 10 observações por nível de tratamento, é pouco provável que termine com esse número. A despeito de seus melhores esforços, os dados podem ser perdidos por uma variedade de razões, incluindo a falha nos equipamentos, os desastres climáticos, as parcelas perdidas, os distúrbios ou erros humanos, a transcrição imprópria de dados e as alterações ambientais. A Regra dos 10 possibilita a coleta de dados com um poder estatístico razoável para revelar padrões.[8] No Capítulo 7, discutiremos delineamentos experimentais eficientes e estratégias para maximizar a quantidade de informação que você poderá extrair de seus dados.

Estudos em larga escala e impactos ambientais

A Regra dos 10 é útil para experimentos manipulativos em pequena escala, nos quais as unidades de estudo (parcelas, folhas, etc.) são de tamanho manejável. Mas ela não se aplica a experimentos de ecossistemas em larga escala, como a manipulação de um lago inteiro, pois as réplicas podem ser indisponíveis ou muito caras. A Regra dos 10 também não se aplica a muitos estudos de impacto ambiental, onde a avaliação de um impacto se faz necessária para um único local. Nesses casos, a melhor estratégia é usar o **delineamento ADCI** (antes-depois; controle-impacto). Em alguns delineamentos ADCI, a replicação é obtida com o passar do tempo: os locais de controle e impactados são amostrados repetidamente, antes e depois do impacto. A falta de replicação espacial por si só restringe as inferências ao local do impacto (que deve ser o foco do estudo) e requer que o impacto não seja confundido com outros fatores que possam co-variar com o impacto. A falta de replicação espacial em delineamentos simples de ADCI é controversa (Underwood, 1994; Murtaugh, 2002b), mas, em muitos casos, são a melhor opção de delineamento (Stewart-Oaten & Bence, 2001), especialmente se usados em modelagem de séries temporais explícitas (Carpenter et al., 1989). Retornaremos ao ADCI e suas alternativas nos Capítulos 7 e 10.

[8] A Regra dos 5 também é útil. Se você quiser estimar a curvatura ou a não linearidade de uma resposta, precisa usar ao menos cinco níveis de variável preditora. Como discutiremos no Capítulo 7, uma solução melhor é usar um delineamento de regressão, no qual a variável preditora é contínua, ao invés de categórica e com um número fixo de níveis.

GARANTINDO INDEPENDÊNCIA

A maioria das análises estatísticas assume que as réplicas são independentes umas das outras. Por **independência** queremos dizer que as observações coletadas em uma réplica não têm influência sobre as observações coletadas em outra. A falta de independência é mais facilmente entendida em um contexto experimental. Suponha que você está estudando a resposta de beija-flores polinizadores à quantidade de néctar produzido por flores. Você estabelece duas parcelas adjacentes de 5 m x 5 m. Uma parcela é um controle; a adjacente é uma parcela com remoção de néctar, da qual você retira todo o néctar das flores. Além disso, você mede a visita dos beija-flores às flores das duas parcelas. Na parcela controle, você mediu, aproximadamente, 10 visitas/hora, comparada a apenas 5 visitas/hora na com remoção.

Contudo, enquanto coletava os dados, percebeu que uma vez que os pássaros chegavam à parcela de remoção, eles partiam imediatamente, e os *mesmos pássaros* visitavam, então, a parcela-controle adjacente (Figura 6.5A). Óbvio, os dois conjuntos de observações não são independentes um do outro. Se as parcelas-controle e de tratamento tivessem sido mais amplamente separadas no espaço, os números obtidos poderiam ser diferentes, e a média de visitas na controle poderia ser de apenas 7 visitas/hora, em vez de 10 visitas/hora (Figura 6.5B). Quando as duas parcelas são adjacentes uma a outra, a falta de independência infla a diferença entre elas, talvez levando a um baixo e enganoso valor de P e a um erro Tipo I (rejeição incorreta de uma hipótese nula verdadeira; *ver* Capítulo 4). Em outros casos, a falta de independência pode diminuir as diferenças aparentes entre tratamentos, contribuindo para um erro Tipo II (aceitar incorretamente uma hipótese nula falsa). Infelizmente, a falta de independência infla ou desinfla tanto o valor de P quanto o poder a graus desconhecidos.

A melhor garantia contra a falta de independência é assegurar que as réplicas dentro e entre tratamentos sejam separadas umas das outras por espaço ou tempo suficientes para garantir que elas não afetem umas às outras. Infelizmente, raras vezes sabemos a distância ou o espaçamento que deve haver, e isso é verdade tanto para estudos experimentais quanto para observacionais. Devemos usar o senso comum e todo o conhecimento biológico quanto possível. Tente olhar o mundo através da perspectiva do organismo para imaginar o quão distante deve separar as amostras. Estudos-piloto, se viáveis, também podem sugerir o espaçamento apropriado para garantir independência.

Então, por que não apenas maximizar a distância ou o tempo entre as amostras? Primeiro, como descrevemos antes, se torna mais caro coletar dados conforme a distância entre as amostras aumenta. Segundo, mover as amostras para muito longe pode introduzir novas fontes de variação, por causa de diferenças (heterogeneidade) dentro ou entre hábitats. Queremos nossas réplicas perto o su-

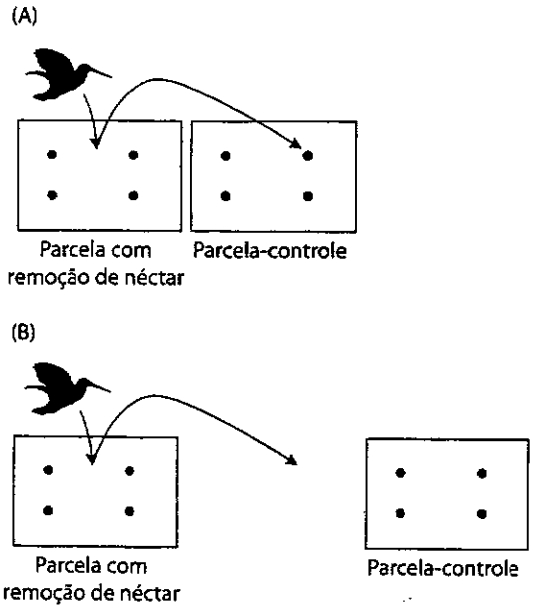


Figura 6.5 O problema de falta de independência em estudos ecológicos é ilustrado por um delineamento experimental em que beija-flores forrageiam em busca de néctar em parcelas controle e em parcelas das quais o néctar foi removido de todas as flores. (A) Em um esboço sem independência, as parcelas com remoção de néctar e controle são adjacentes uma a outra e beija-flores que entram na parcela de remoção de néctar imediatamente a deixam e começam a forragear na parcela-controle adjacente. Como consequência, os dados coletados na parcela controle não são independentes dos dados coletados na parcela com remoção de néctar: as respostas em um tratamento influenciam as respostas do outro. (B) Se o esboço for modificado de forma que as parcelas sejam amplamente separadas, os beija-flores que deixarem a parcela com remoção de néctar não necessariamente entrarão na parcela-controle. As duas parcelas são independentes e os dados coletados em uma parcela não serão influenciados pela presença da outra parcela. Apesar de ser fácil ilustrar o problema potencial da falta de independência, na prática pode ser muito difícil saber por antecipação a escala espacial e temporal que garantirá independência estatística.

ficiente para garantir que estamos amostrando condições relativamente homogêneas ou consistentes, mas distantes o suficiente para garantir que as respostas que medirmos são independentes umas das outras.

Apesar de sua importância central, o problema da independência quase nunca é discutido de forma explícita em artigos científicos. Na seção dos métodos de um artigo, você provavelmente lerá uma sentença como: "Medimos 100 plântulas, selecionadas aleatoriamente, crescendo sob luz solar total. Cada plântula medida estava a pelo menos 50 cm de sua vizinha mais próxima." O que os autores querem dizer é: "Não sabemos quão longe as observações deveriam estar umas das outras para garantir independência. Contudo, 50 cm nos pareceu

razoável para as pequenas plântulas que estudamos. Se tivéssemos escolhido distâncias maiores que 50 cm, poderíamos não ter coletado todos os dados sob luz solar total, e algumas poderiam ter sido coletadas na sombra, o que, obviamente, influenciaria nossos resultados".

EVITANDO FATORES CONFUNDIDOS

Quando os fatores se confundem uns com os outros, seus efeitos não podem ser facilmente esclarecidos. Retornando ao exemplo dos beija-flores, suponha que prudentemente separamos as parcelas-controle e com remoção de néctar, mas, sem querer, colocamos a parcela com remoção em uma encosta ensolarada e a parcela-controle em um vale frio (Figura 6.6). Os beija-flores forrageiam com menos frequência na parcela com remoção (7 visitas/hora), e as duas parcelas agora são distantes o suficiente para garantir que não há problemas com independência. Contudo, os beija-flores naturalmente tendem a evitar o forrageio no vale frio, de modo que a taxa de forrageamento também é baixa nessa parcela (6 visitas/hora). Como os tratamentos se confundem com diferenças de temperatura, não podemos separar os efeitos de preferências de forrageamento das preferências termais. Nesse caso, as duas forças se anulam, levando as taxas de forrageamento comparáveis entre as duas parcelas, embora por razões muito diferentes.

Tal exemplo pode parecer um pouco forçado. Sabendo a preferência termal de beija-flores, não montaríamos tal experimento. O problema é que provavelmente existam variáveis não medidas ou desconhecidas – mesmo em um ambiente aparentemente homogêneo – que podem ter efeitos igualmente fortes em nosso experimento. E, se estivermos conduzindo um experimento natural, emperrare-

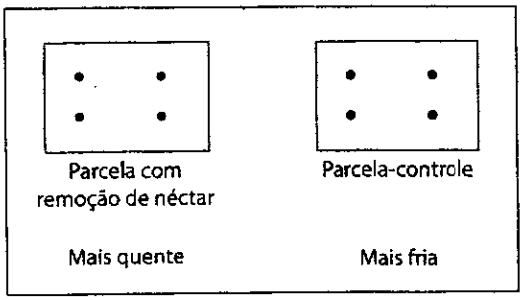


Figura 6.6 Um delineamento experimental confundido. Como na Figura 6.5, o estudo estabelece parcelas-controle e com remoção de néctar para avaliar as respostas no forrageamento de beija-flores. Nesse delineamento, embora as parcelas sejam suficientemente afastadas para garantir independência, foram colocadas em pontos diferentes ao longo de um gradiente termal. Por consequência, os efeitos dos tratamentos são confundidos com diferenças de temperatura. O resultado líquido é que o experimento compara dados de uma parcela quente com remoção de néctar com dados de uma parcela-controle fria.

mos em quaisquer fatores confundidos presentes no ambiente. Em um estudo observacional do forrageamento de beija-flores, poderemos não ser aptos para encontrar parcelas que difiram apenas em seus níveis de recompensa de néctar, mas que também não difiram em temperatura ou outros fatores que sabidamente afetam o comportamento de forrageamento.

REPLICAÇÃO E RANDOMIZAÇÃO

A dupla ameaça dos fatores confundidos e da falta de independência parece ameaçar todas as nossas conclusões estatísticas e ainda trazer suspeitas aos nossos experimentos. Incorporar a replicação e a randomização aos delineamentos experimentais pode compensar fortemente os problemas introduzidos por fatores confundidos e por falta de independência. Por replicação, queremos nos referir ao estabelecimento de múltiplas parcelas ou às observações dentro do mesmo tratamento ou grupo comparado. Por randomização, referimo-nos à atribuição aleatória de tratamentos ou à seleção de amostras.[9]

Vamos retornar mais uma vez ao exemplo dos beija-flores. Se seguirmos os princípios da randomização e da replicação, estabeleceremos muitas réplicas de parcelas-controle e com remoção de néctar (idealmente, no mínimo 10 de cada). A localização de cada uma dessas parcelas na área de estudo é aleatória, e a atribuição do tratamento (controle ou com remoção) a cada parcela também será aleatória (Figura 6.7).[10]

Como a replicação e a randomização podem reduzir o problema de fatores confundidos? Tanto a encosta quente, o vale frio, quanto os diversos locais intermediários terão parcelas múltiplas de controles e de tratamentos com remoção de néctar. Logo, o fator temperatura não estará mais confundido com o tratamento, como todos os tratamentos ocorrem dentro de cada nível de temperatura. Como benefício adicional, esse delineamento também permitirá o teste dos efeitos da temperatura como uma covariável sobre comportamento de forrageamento dos

[9] Muitas amostragens classificadas como aleatórias na verdade são casuais. Uma amostragem verdadeiramente aleatória significa usar um gerador de números aleatórios (como o arremesso de uma moeda honesta, o lançamento de um dado honesto ou o uso de um algoritmo de computador confiável para produzir números aleatórios) para decidir qual réplica usar. Em contraste, com amostragem casual, um ecólogo segue um conjunto de critérios gerais [p. ex., árvores adultas têm um diâmetro à altura do peito de mais de 3 cm (dap = 1,3 m)], ou selecionam locais ou organismos espaçados homogeneamente, ou por conveniência, dentro da área amostral. Muitas vezes é necessário amostrar casualmente em algum nível, porque a amostragem aleatória não é eficiente para muitos tipos de organismos, em especial se sua distribuição é espacialmente agregada, formando manchas. Contudo, uma vez que um conjunto de organismos ou de locais é identificado, a randomização deve ser usada para amostrar ou para atribuir os diferentes grupos de tratamentos às réplicas.

[10] A randomização leva algum tempo, e você deve randomizar o máximo possível antes de ir ao campo. É fácil gerar números aleatórios e simular amostragem aleatória com planilhas de computador. Porém, muitas vezes será o caso de você precisar gerar números aleatórios em campo. Moedas e dados (especialmente dados de 10 faces) são úteis para esse propósito. Um estratagema

(*Continua*)

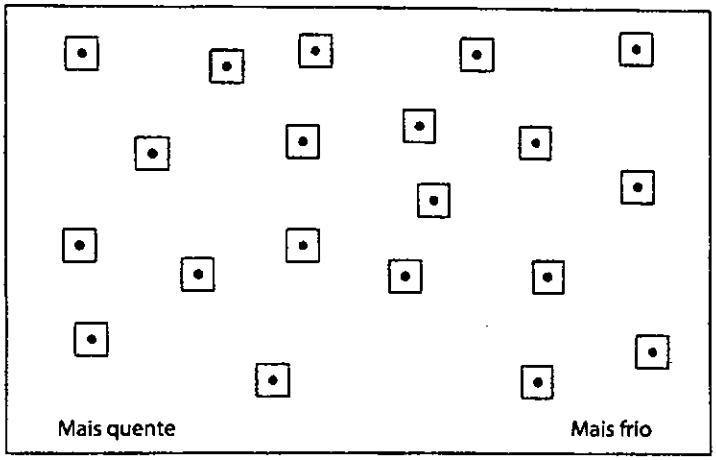


Figura 6.7 Um delineamento experimental devidamente replicado e aleatorizado. O estudo estabelece parcelas como na Figura 6.6. Cada quadrado representa uma réplica da parcela-controle (pontos pretos) ou da parcela com remoção de néctar (pontos cinzas). As parcelas são separadas por distância suficiente para garantir independência, mas sua localização dentro do gradiente de temperatura foi aleatorizada. Há 10 réplicas para cada um dos dois tratamentos. A escala espacial desse desenho é maior que na Figura 6.6.

beija-flores, independente dos níveis de néctar (*ver* Capítulos 7 e 10). É verdade que a visitação dos beija-flores ainda será mais frequente na encosta quente que no vale frio, mas isso será verdadeiro tanto para as réplicas quanto para as com remoção de néctar. A temperatura irá adicionar mais variação aos dados, mas isso não conduzirá os resultados, porque as parcelas quentes e as frias estarão aproximadamente distribuídas em igualdade entre os tratamentos controle e com remoção. É claro, se soubéssemos por antecipação que a temperatura seria um importante determinante do comportamento de forrageamento, poderíamos não ter usado esse delineamento para o experimento. A randomização minimiza a

[10] (*Continuação*)

esperto é usar um conjunto de moedas como um gerador de números aleatórios binários. Por exemplo, suponha que você deverá atribuir a cada uma de suas réplicas um tratamento dentre 8 possibilidades, e você deseja fazer isso de maneira aleatória. Arremesse 3 moedas e converta o padrão de caras e coroas em um número binário (i. e., um número na base 2). Assim, a primeira moeda indica os 1(s), a segunda moeda indica os 2(s) e a terceira moeda indica os 4(s), e assim por diante. O arremesso de 3 moedas resultará um número aleatório entre 0 e 7. Se os três arremessos derem cara, coroa, cara (CACOCA), você terá um 1 no lugar do 1, um 0 no lugar do 2, e 1 no lugar 4. Seu número aleatório será 1 + 0 + 4 = 5. Um lançamento (COCACO) dará 0 + 2 + 0 = 2. Três coroas dariam 0 (0 +0 + 0) e três caras dariam 7 (1 + 2 + 4). Arremessar 4 moedas resultará em 16 números inteiros e 5 moedas, 32.

Um método ainda mais simples é levar um cronômetro digital para o campo. Deixe o cronômetro correr por alguns segundos e pare-o sem olhar para ele. O último dígito que mede a hora em centésimos de segundo pode ser usado como um dígito aleatório uniforme de 0 a 9. A análise estatística de 100 desses dígitos aleatórios passou em todos os testes diagnósticos padrão para testar aleatoriedade e uniformidade (B. Inouye, comunicação pessoal).

confusão dos tratamentos com variáveis desconhecidas, ou não medidas, na área de estudo.

É menos óbvio como a randomização e a replicação reduzem o problema de falta de independência entre amostras. Afinal de contas, se as parcelas são muito próximas umas das outras, as visitas de forrageamento não serão independentes, não dependendo da quantidade de réplicas ou de randomização. Sempre que possível, devemos usar o senso comum e o conhecimento de biologia para separar parcelas ou amostras por uma distância ou um intervalo amostral mínimo, a fim de evitar dependência. Contudo, se não conhecemos todas as forças que causam a dependência, a disposição aleatória das parcelas acima de uma determinada distância mínima irá garantir que o espaçamento entre as parcelas seja variável. Algumas parcelas ficarão relativamente próximas e outras, longe. Portanto, o efeito da dependência será forte entre alguns pares de parcelas, fraco entre outros e não existirá entre o restante. Tais efeitos variáveis podem cancelar um ao outro e podem reduzir as chances de que os resultados sejam consistentemente conduzidos pela falta de independência.

Por fim, note que a randomização e a replicação somente são efetivas quando usadas em conjunto. Se não replicarmos, mas simplesmente atribuirmos de forma aleatória as parcelas-controle e de tratamento para a encosta ou para o vale, o delineamento ainda assim será confundido (Figura 6.6). Similarmente, se replicarmos o delineamento, mas atribuirmos todos os 10 controles no vale a todas as 10 réplicas com remoção na encosta, o delineamento também será confundido (Figura 6.8). É somente quando usamos múltiplas parcelas e atribuímos os tratamentos aleatoriamente que o efeito confundido da temperatura é removido do delineamento (Figura 6.7). De fato, é justo dizer que qualquer delineamento sem réplicas sempre vai ser confundido com um ou mais fatores ambientais.[11]

Apesar do conceito de randomização ser simples, ele precisa ser aplicado em vários estágios do delineamento. Primeiro, a randomização se aplica somente a um espaço amostral bem-definido e inicialmente não aleatório. O espaço amostral não significa simplesmente a área física da qual as réplicas são amostradas (embora este seja um aspecto importante do espaço amostral). Em vez disso, o espaço amostral se refere a um conjunto de elementos que possuem condições similares, não idênticas.

[11] Apesar de fatores confundidos serem fáceis de reconhecer em um experimento de campo desse tipo, não quer dizer que o mesmo problema exista em experimentos de laboratório ou em casas de vegetação (estufas). Se nós criamos larvas de insetos em temperaturas altas e baixas em duas câmaras ambientais, esse será um delineamento confundido, porque todas as larvas de temperatura alta estarão em uma câmara e todas as de temperatura baixa estarão em outra. Se outros fatores ambientais, que não a temperatura, também difiram entre as câmaras, seus efeitos estarão confundidos com a temperatura. A solução correta seria criar cada larva em câmaras individuais, garantindo, desse modo, que cada réplica seja verdadeiramente independente e que a temperatura não seja confundida com outros fatores. Mas esse tipo de delineamento é muito oneroso e é um desperdício de espaço empregá-lo. Talvez se possa argumentar que as câmaras ambientais ou casas de vegetação realmente diferem apenas em temperatura e em nenhum outro fator, mas esse é apenas um pressuposto que deve ser testado de maneira explícita. Em muitos casos, o ambiente em câmaras ambientais é surpreendentemente heterogêneo, tanto dentro quanto entre câmaras. Potvin (2001) discute como esta variabilidade pode ser medida e então utilizada para delinear melhor experimentos de laboratório.

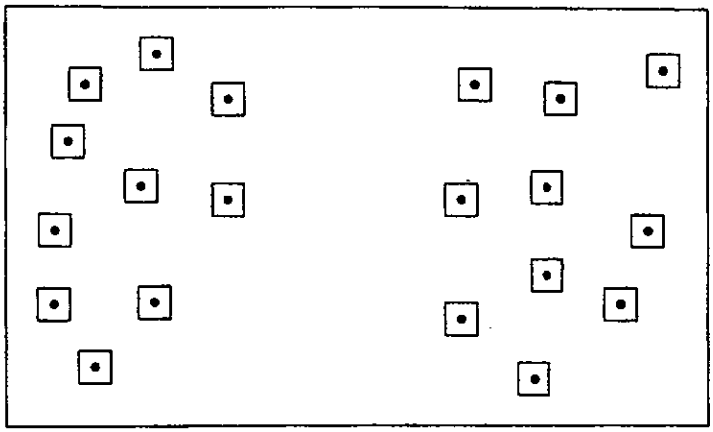


Figura 6.8 Um delineamento replicado, porém confuso. Como nas Figuras 6.5, 6.6 e 6.7, o estudo estabelece parcelas experimentais de controle e com remoção de néctar para avaliar as respostas no forrageamento de beija-flores. Cada quadrado representa uma parcela-controle (pontos pretos) ou parcelas com remoção de néctar (pontos cinza). Se os tratamentos forem replicados, mas não atribuídos aleatoriamente, o delineamento ainda confunde tratamento com gradientes ambientais adjacentes. A replicação combinada com randomização e espaçamento suficiente entre as réplicas (Figura 6.7) é a única garantia contra a falta de independência (Figura 6.5) e fatores confundidos (Figuras 6.6 e 6.8).

Exemplos de um espaço amostral podem incluir indivíduos de trutas que estão reprodutivamente maduros, clareiras criadas por incêndios, antigos campos abandonados 10-20 anos atrás, ou grandes corais branqueados* de 5 a10 metros de profundidade. Se esse espaço amostral foi definido claramente, locais, indivíduos ou réplicas que cumpram o critério devem ser escolhidos de maneira aleatória. Como notamos no Capítulo 1, as fronteiras espaciais e temporais do estudo ditarão não apenas o esforço amostral envolvido, mas também o domínio de inferência para as conclusões do estudo.

Uma vez que os locais ou as amostras são selecionados aleatoriamente, os tratamentos devem ser atribuídos a eles sem obedecer ordens ou colocações, garantindo que tratamentos diferentes não fiquem agrupados no espaço ou sejam confundidos com variáveis ambientais.[12] Tantos as amostras quanto os tra-

[12] Se o número de amostras é muito pequeno, mesmo a atribuição aleatória pode levar a uma agregação espacial dos tratamentos. Uma solução poderia ser estabelecer os tratamentos em ordem repetida (...123123...), o que garante a não agregação. Porém, se há falta de independência entre tratamentos, esse delineamento pode exagerar seus efeitos, porque o tratamento 2 sempre irá ocorrer espacialmente entre os de número 1 e 3. Uma solução melhor seria repetir a randomização e, então, testar estatisticamente a conformação para garantir que não haja agregação. *Ver* Hurlbert (1984) para uma discussão aprofundada dos numerosos riscos que podem surgir ao falhar em replicar e randomizar adequadamente experimentos ecológicos.

* N. de T. O branqueamento (do inglês *bleaching*) é a perda dos organismos fotossimbiontes (zooxantelas) presentes nos tecidos do coral. O tecido vivo dos corais sem as algas é translúcido, e por isso o esqueleto feito de carbonato de cálcio fica exposto, resultando no coral branqueado.

tamentos devem ser coletados e aplicados, respectivamente, de forma aleatória. Dessa forma, se as condições ambientais mudarem durante o experimento, os resultados não serão confundidos. Por exemplo, se você faz o censo de os seus controles primeiro, e seu trabalho de campo é interrompido por uma violenta tempestade, quaisquer impactos do mau tempo serão confundidos com suas manipulações, pois o censo em todas as parcelas de tratamento será feito depois da tempestade. Essas mesmas ressalvas se aplicam a estudos não experimentais, nos quais será feito o censo em diferentes parcelas ou locais. A advertência é que fazer o censo aleatório apenas dessa forma pode ser muito ineficiente, porque você em geral não visitará locais próximos em uma ordem consecutiva. Você pode ter de fazer um compromisso entre a randomização estrita e as restrições impostas pela eficiência amostral.

Todos os métodos de análises estatísticas – sejam eles paramétricos, Monte Carlo, ou bayesianos (ver Capítulo 5) – recaem sobre o pressuposto de amostragem aleatória em uma escala espacial ou temporal apropriada. Você deve adquirir o hábito de empregar a randomização sempre que possível em seu trabalho.

DELINEANDO EXPERIMENTOS DE CAMPO E ESTUDOS AMOSTRAIS EFETIVOS

Aqui, estão algumas questões a serem feitas quando delineamos experimentos de campo e estudos amostrais. Embora algumas pareçam ser específicas a experimentos manipulativos, elas também são relevantes a certos experimentos naturais, onde os "controles" podem consistir em parcelas sem uma espécie em particular ou sem um conjunto de condições abióticas.

As parcelas ou os cercados são grandes o suficiente para garantir resultados realísticos?

Experimentos de campo que buscam controlar a densidade de animais devem necessariamente restringir o seu movimento. Se os cercados forem muito pequenos, o movimento, o forrageamento e os comportamentos reprodutivos dos animais podem ser falsos e, por conseguinte, os resultados obtidos não serão interpretáveis (MacNally, 2000a). Tente usar a maior parcela, ou gaiola, viável e apropriada para o organismo que esteja estudando. As mesmas considerações se aplicam aos estudos amostrais: as parcelas precisam ser grandes o suficiente e amostradas em uma escala espacial apropriada para responder sua questão.

Qual é o grão e a extensão do estudo?

Embora muita importância tenha sido depositada na escala espacial de um experimento ou de um estudo amostral, existem, na verdade, dois componentes da escala espacial que precisam ser discutidos: o grão e a extensão. O **grão** é o tamanho da menor unidade do estudo, que geralmente será o tamanho de uma réplica ou parcela individual. A **extensão** é a área total englobada por todas as unidades amostrais no estudo. O grão e a extensão tanto podem ser grandes quanto pequenos (Figura 6.9). Não há uma combinação única de grão e extensão que seja

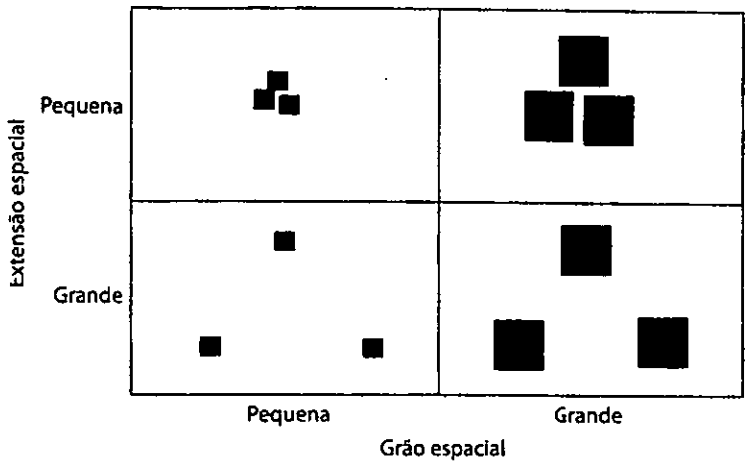


Figura 6.9 Grão espacial e extensão espacial em estudos ecológicos. Cada quadrado representa uma única parcela. O grão espacial mede o tamanho das unidades amostrais, representada por quadrados pequenos ou grandes. A extensão espacial mede a área que engloba todas as réplicas do estudo, representada por quadrados agrupados bem próximos ou amplamente espaçados.

necessariamente correta. Contudo, estudos ecológicos com pequena granulação e pequena extensão, como a captura de besouros com armadilhas do tipo *pitfall* em uma única parcela de floresta, podem, por vezes, ter um escopo muito limitado para permitir conclusões amplas. Por outro lado, estudos com um grande grão, e pequena extensão, como a manipulação de lagos inteiros em um único vale, podem ser bastante informativos. Nossa preferência é por estudos com um grão pequeno, mas extensão média ou grande, como os censos de formigas e plantas em pequenas parcelas (5 m × 5 m) ao longo de todo o estado da Nova Inglaterra [*New England*] (Gotelli & Ellison, 2002a,b) ou da América do Norte oriental (Gotelli & Arnett, 2000), ou em pequenas ilhas de mangue no Caribe (Farnsworth & Ellison, 1996a). O grão pequeno permite manipulações experimentais e observações feitas em escalas que são relevantes para o organismo, mas a larga extensão expande o domínio de inferência dos resultados. Ao determinar o grão e a extensão, você deve considerar tanto a questão que está tentando responder como as restrições sobre sua amostragem.

A amplitude dos tratamentos ou das categorias dos censos engloba ou abrange a amplitude de condições ambientais possíveis?

Muitos experimentos de campo descrevem suas manipulações como "englobando ou abrangendo a gama de condições encontradas no campo." Contudo, se você está tentando modelar uma mudança climática, ou ambientes alterados, pode ser necessário incluir, também, condições que estão fora da amplitude daquelas normalmente encontradas em campo.

Foram estabelecidos controles apropriados para garantir que os resultados refletem variação apenas no fator de interesse?

É raro que uma manipulação irá mudar um, e apenas um fator por vez. Por exemplo, se você cercar plantas com uma gaiola para excluir herbívoros, você também terá alterado o sombreamento e o regime de umidade. Se você simplesmente comparar essas plantas a controles sem manipulação, os efeitos dos herbívoros estarão confundidos com as diferenças de sombreamento e umidade. O erro mais comum em delineamentos experimentais é estabelecer um conjunto de parcelas sem manipulação e tratá-las como controle. Em geral, um conjunto adicional de parcelas-controle, que contenham um mínimo de alteração, será necessário para controlar de maneira adequada as manipulações. No exemplo retromencionado, uma gaiola com um dos lados aberto permitirá o acesso de herbívoros, mas ainda incluirá os efeitos do sombreamento da gaiola. Com esse simples delineamento de três tratamentos (sem manipulação, gaiola controle, exclusão de herbívoros), você poderá fazer os seguintes contrastes:

1. *Sem manipulação* versus *Gaiola controle*. Esta comparação revela a extensão na qual o sombreamento e as mudanças físicas devido à gaiola *per se* estão afetando o crescimento e respostas das plantas.
2. *Gaiola controle* versus *Exclusão de herbívoros*. Esta comparação revela a extensão na qual a herbivoria afeta o crescimento das plantas. Tanto as parcelas controle quanto as de exclusão de herbívoros sofrem os efeitos de sombreamento da gaiola. Assim, qualquer diferença entre elas pode ser atribuída ao efeito dos herbívoros.
3. *Sem manipulação* versus *Exclusão de herbívoros*. Esta comparação mensura o efeito combinado dos herbívoros com o do sombreamento no crescimento das plantas. Como o experimento é delineado para medir apenas o efeito dos herbívoros, esta comparação em particular confunde entre os efeitos dos tratamentos com os gerados pelo uso da gaiola.

No Capítulo 10, explicaremos como você pode usar contrastes após uma análise de variância para quantificar essas comparações.

Todas as réplicas foram manipuladas da mesma forma, exceto para a aplicação intencional do tratamento?

Novamente, controles apropriados em geral requerem mais que ausência de manipulação. Se você precisar "empurrar" as plantas para aplicar o tratamento, deve fazer o mesmo com as plantas das parcelas-controle (Salisbury, 1963, Jaffe, 1980). Em um experimento de transplante recíproco de larvas de insetos, animais vivos podem ser enviados via correio para locais distantes e inseridos em novas populações, no campo. O controle apropriado é um conjunto de animais que são reinseridos nas populações das quais eles foram coletados. Esses animais também deverão receber o mesmo "tratamento postal", sendo enviados pelo sistema de correios para garantir que eles recebam a mesma carga de estresse que aqueles que foram transplantados para locais distantes. Se você não for cuida-

doso de garantir que todos os organismos foram tratados identicamente em seu experimento, seus tratamentos serão confundidos com os efeitos da manipulação (Cahill et al., 2000).

As covariáveis apropriadas foram medidas em cada réplica?

As covariáveis são variáveis contínuas (*ver* Capítulo 7) que afetam muito a variável resposta, mas não são necessariamente controladas ou manipuladas pelo pesquisador. Exemplos incluem as variações de temperatura, de sombreamento, de pH ou de densidade de herbívoros entre parcelas. Diferentes métodos estatísticos, como a análise de covariância (*ver* Capítulo 10), podem ser usados para quantificar o efeito de covariáveis.

Contudo, você deve evitar a tentação de medir todas as covariáveis concebíveis em uma parcela apenas porque tem a instrumentação (e o tempo) para fazê-lo. Você rapidamente terminará com um conjunto de dados em que terá mais variáveis medidas do que tem de réplicas, causando problemas adicionais na análise (Burnham & Anderson, 2002). É melhor escolher por antecipação as covariáveis mais relevantes biologicamente, mensurar apenas aquelas covariáveis e usar a replicação suficiente. Lembre-se também, de que a mensuração de covariáveis também é útil, mas não é um substituto para a randomização e replicação adequadas.

RESUMO

O bom delineamento de um experimento em ecologia requer primeiro o claro estabelecimento de uma questão a ser respondida. Tanto experimentos manipulativos como observacionais podem responder perguntas ecológicas, e cada tipo de experimento tem suas próprias forças e fraquezas. Pesquisadores devem considerar a adequação de utilizar um experimento de pressão *versus* um de pulso, e se a replicação será no espaço (experimentos fotográficos), no tempo (experimentos de trajetória), ou em ambos. A falta de independência e fatores confundidos podem comprometer a análise estatística dos dados, tanto de estudos manipulativos, quanto observacionais. A randomização, a replicação e o conhecimento de ecologia e da história natural dos organismos são as melhores garantias contra a falta de independência e os fatores confundidos. Sempre que possível, tente usar no mínimo 10 observações por grupo de tratamentos. Experimentos de campo geralmente requerem controles delineados com cautela, levando em conta efeitos de manipulação e outras alterações não intencionais. A medida de covariáveis ambientais apropriadas podem ser utilizadas para explicar a variação não controlada, embora ela não seja um substituto para a randomização e para a replicação.

7 Um Bestiário de Delineamentos Experimentais e Amostrais

Em um estudo experimental, precisamos decidir sobre um conjunto de manipulações biologicamente realistas que incluam controles apropriados. Em um estudo observacional, temos de decidir quais variáveis medir para melhor responder às questões que descobrimos. Essas decisões são muito importantes e foram o assunto do Capítulo 6. Nesse capítulo discutimos, ainda, delineamentos específicos para estudos experimentais e amostrais em ecologia e ciência ambiental. O delineamento de um experimento ou estudo observacional se refere a como as réplicas são fisicamente organizadas no espaço e como são amostradas com o passar do tempo. O delineamento do experimento está intimamente ligado aos detalhes de replicação, randomização e independência (*ver* Capítulo 6). Certos tipos de delineamentos têm se revelado muito robustos para a interpretação e análise de dados de campo. Outros são mais difíceis para analisar e interpretar. No entanto, você não consegue tirar leite de pedra, e mesmo as análises estatísticas mais sofisticadas não podem salvar um delineamento ruim.

Primeiro, apresentamos um esquema de trabalho simples para classificar os delineamentos de acordo com os tipos de variáveis, se dependentes ou independentes. Depois, descrevemos um pequeno número de delineamentos úteis em cada categoria. Discutimos cada um e os tipos de questões que eles podem ser usados para resolver, os ilustramos com um conjunto de dados simples e descrevemos as vantagens e desvantagens de cada caso. Os detalhes de como analisar dados a partir desses delineamentos estão nos Capítulos 9 a 12.

---

* N. de R.T. Ver nota da p. XI.

A literatura sobre delineamentos experimentais e amostrais é vasta (p. ex., Cochran & Cox, 1957; Winer, 1991; Underwood, 1997; Quinn & Keough, 2002) e aqui apresentamos apenas uma seleta cobertura. Nos restringimos aos delineamentos que são práticos e úteis para ecólogos e cientistas ambientais e que tenham se revelado bem-sucedidos em estudos de campo.

VARIÁVEIS CATEGÓRICAS *VERSUS* VARIÁVEIS CONTÍNUAS

Primeiro distinguiremos **variáveis categóricas** e **variáveis contínuas.** As variáveis categóricas são classificadas em uma entre duas ou mais categorias. Exemplos ecológicos incluem sexo (macho, fêmea), categoria trófica (produtor, herbívoro, carnívoro) e tipo de hábitat (sombra, sol). As variáveis contínuas são medidas em uma escala numérica contínua, podendo assumir uma gama de números reais ou valores inteiros. Exemplos incluem as medidas do tamanho de indivíduos, a riqueza de espécies, a cobertura de hábitats e a densidade populacional.

Muitos textos de estatística realizam uma distinção adicional entre variáveis puramente categóricas, nas quais as categorias não são variáveis ordenadas, e "ranqueadas" (ou ordinais), cujas categorias são ordenadas com base em uma escala numérica. Um exemplo de variável ordinal pode ser um escore numérico (0, 1, 2, 3 ou 4) atribuído à quantidade de luz solar que atinge o solo da floresta: 0 para 0-5% de luz; 1 para 6-25%; 2 para 26-50%; 3 para 51-75%; e 4 para 76-100%. Em muitos casos, os métodos utilizados para analisar dados contínuos também podem ser aplicados a dados ordinais. Em poucos, no entanto, os dados ordinais são melhor analisados com métodos de Monte Carlo, discutidos no Capítulo 5. Neste livro, usamos o termo *variável categórica* para nos referirmos tanto a variáveis categóricas ordenadas quanto às não ordenadas.

A distinção entre variáveis contínuas e categóricas nem sempre é óbvia: em vários casos, a designação depende simplesmente de como o investigador escolheu medir a variável. Por exemplo, uma variável categórica de hábitat como sol/sombra poderia ser medida, em uma escala contínua, usando um medidor de luz e registrando a intensidade luminosa em diferentes lugares. Ao contrário, uma variável contínua como a salinidade poderia ser classificada em três níveis (baixo, médio, e alto) e tratada como uma variável categórica. Reconhecer o tipo de variável que você está medindo é importante, pois diferentes delineamentos se baseiam em variáveis categóricas e em contínuas.

No Capítulo 2, distinguimos dois tipos de variáveis aleatórias: as discretas e as contínuas. Qual a diferença entre variáveis aleatórias discretas ou contínuas por um lado e entre variáveis categóricas e contínuas por outro? Variáveis aleatórias discretas ou contínuas são funções matemáticas para gerar valores associados com distribuições de probabilidade. Em contraste, variáveis categóricas e contínuas descrevem os tipos de dados que nós realmente medimos em campo ou em laboratório. Variáveis contínuas em geral podem ser modeladas como variáveis aleatórias contínuas, enquanto tanto as categóricas quanto as ordinais geralmente podem ser modeladas como variáveis aleatórias discretas. Por exemplo, a variável

categórica *sexo* pode ser modelada como uma variável aleatória binomial; a variável numérica *altura* pode ser modelada como aleatória normal; e a variável ordinal *luz atingindo o solo da floresta* pode ser modelada como aleatória binomial, de Poisson, ou uniforme.

VARIÁVEIS DEPENDENTES E INDEPENDENTES

Após identificar os tipos de variáveis com as quais você está trabalhando, o próximo passo é designar as **variáveis dependentes** e as **independentes**. Esta atribuição implica uma hipótese de causa e efeito que você está tentando testar. A do tipo dependente é a **variável resposta**, que você está medindo e para a qual você está tentando determinar as causas. Em um gráfico de dispersão de duas variáveis, a dependente ou resposta é chamada de variável *Y*, e geralmente é plotada na **ordenada** (eixo vertical ou *y*). A do tipo independente é a **variável preditora**, que você supõe ser responsável pela alteração na variável resposta. No mesmo gráfico de dispersão das duas variáveis, a independente ou preditora é chamada de variável *X*, e em geral é plotada na **abscissa** (eixo horizontal ou *x*).[1]

Em um estudo experimental, você com frequência manipula ou controla diretamente os níveis da independente e mede a resposta da dependente. Em um estudo observacional, você depende da variação natural da variável independente entre as réplicas. Tanto em estudos naturais quanto em observacionais, você não sabe de antemão o poder da variável preditora. De fato, testa com frequência a hipótese nula estatística de que a instabilidade na variável resposta não é relacionada com a da variável preditora, bem como não é maior que o esperado ao acaso ou por erro amostral. A hipótese alternativa é que o acaso não pode responder completamente por essa variação e que ao menos parte dela pode ser atribuída à variável preditora. Você também pode estar interessado em estimar a magnitude do efeito do preditor ou variável causal sobre a variável resposta.

QUATRO CLASSES DE DELINEAMENTOS EXPERIMENTAIS

Combinando os tipos de variáveis – categórica *versus* contínua, dependente *versus* independente – obtemos quatro classes diferentes de delineamentos (Tabela 7.1). Quando as variáveis independentes forem contínuas, as classes são regressão (variáveis dependentes contínuas) ou regressão logística (variáveis dependentes categóricas). Quando as variáveis independentes são categóricas, as classes são ANOVA (variável dependente contínua) ou tabular (variável dependente categórica). Nem todos os delineamentos se ajustam de maneira satisfatória nessas quatro categorias. A análise de covariância (ANCOVA) é usada quando existem

[1] Naturalmente, apenas plotar uma variável no eixo-*x* não é uma garantia de que ela realmente seja a variável preditora. Em experimentos naturais, em especial, a direção de causa e efeito nem sempre é clara, mesmo que as variáveis medidas possam ser altamente correlacionadas (*ver* Capítulo 6).

TABELA 7.1 Quatro classes de delineamentos experimentais e amostrais

	Variável independente	
Variável dependente	Contínua	Categórica
Contínua	Regressão	ANOVA
Categórica	Regressão logística	Tabular

Diferentes tipos de delineamentos são usados dependendo de quando as variáveis independentes e dependentes são contínuas ou categóricas. Quando ambas forem contínuas, um delineamento de regressão é usado. Se a variável dependente é categórica e a variável independente é contínua, emprega-se um delineamento de regressão logística. A análise de delineamentos de regressão é tratada no Capítulo 9. Se a variável independente é categórica e a variável dependente é contínua, usa-se um delineamento de análise de variância (ANOVA). A análise de delineamentos de ANOVA é descrita no Capítulo 10. Por fim, se as variáveis dependentes e independente forem categóricas, usa-se, então, um delineamento tabular. A análise de dados tabulares é descrita no Capítulo 12.

duas variáveis independentes, uma que é categórica e outra que é contínua (a covariável). A ANCOVA é discutida no Capítulo 10. A Tabela 7.1 categoriza dados univariados, nos quais existe uma única variável dependente. Se, em vez disso, tivermos um vetor de variáveis dependentes correlacionadas, contamos com a análise de variância multivariada (MANOVA) ou outros métodos multivariados descritos no Capítulo 12.

Delineamentos de regressão

Quando as variáveis independentes são medidas em escalas numéricas contínuas (*ver* Figura 6.1 para um exemplo), o esquema amostral é um **delineamento de regressão**. Se a variável dependente também é medida em uma escala contínua, usamos modelos de regressão linear ou não linear para analisar os dados. Se é medida em uma escala ordinal (uma resposta ordenada), usamos regressão logística. Estes três tipos de modelos de regressão são discutidos em detalhes no Capítulo 9.

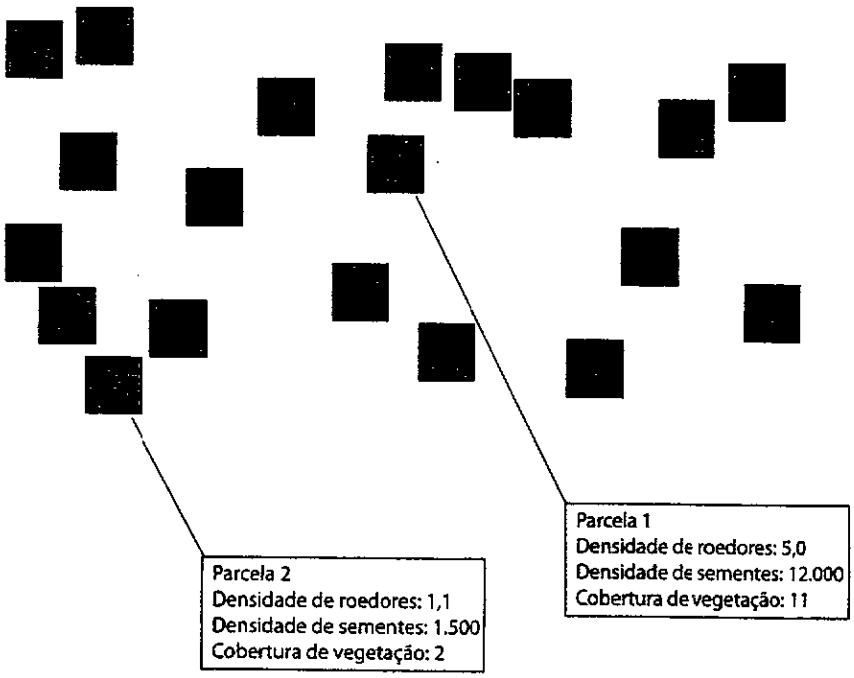
REGRESSÃO DE UM FATOR Um delineamento de regressão é simples e intuitivo. Colete dados em um conjunto de réplicas independentes. Para cada réplica, meça as variáveis preditora e a resposta. Em um estudo observacional, nenhuma das duas variáveis é manipulada, e sua amostragem é ditada pelos níveis de variação natural da variável independente. Por exemplo, suponha que sua hipótese seja de que a densidade de roedores do deserto seja controlada pela disponibilidade de sementes (Brown & Leiberman, 1973). Você poderia amostrar 20 parcelas independentes, cada uma escolhida para representar um nível diferente de abundância de sementes. Em cada parcela, você mede a densidade de sementes e a densidade de roedores do deserto (Figura 7.1). Os dados são organizados em uma planilha, na qual cada linha é uma parcela diferente, e cada coluna é uma variável preditora ou resposta diferente. Os valores em cada linha representam as medidas obtidas em uma única parcela.

Em um estudo experimental, os níveis da variável preditora são controlados e manipulados diretamente, e você mede a variável resposta. Como sua hipótese é de que a densidade de sementes é responsável pela de roedores do deserto (e não o contrário), você poderia manipular a densidade de sementes em um estudo experimental, seja adicionando ou removendo sementes para alterar sua disponibilidade para os roedores. Tanto no estudo experimental quanto no observacional, seu pressuposto é de que a variável preditora é uma causal: mudanças no valor do preditor (densidade de sementes) poderiam *causar* uma alteração no valor da variável resposta (densidade de roedores). Isto é muito diferente de um estudo em que você poderia examinar a correlação (covariância estatística) entre duas variáveis. A correlação não especifica uma relação de causa e efeito entre as duas variáveis.[2]

Em adição às ressalvas usuais de replicação adequada e independência dos dados (Capítulo 6), dois princípios devem ser seguidos ao delinear um estudo de regressão:

1. *Garanta que a amplitude dos valores amostrados para a variável preditora capture toda a gama de respostas da variável resposta.* Se a preditora for amostrada em uma amplitude muito limitada, pode surgir uma relação estatística fraca ou inexistente entre as variáveis preditora e resposta, mesmo que elas sejam relacionadas (Figura 7.2). Uma gama de amostragem limitada torna o estudo suscetível ao erro estatístico do Tipo II (falha em rejeitar uma hipótese nula falsa; *ver* Capítulo 4).
2. *Garanta que a distribuição dos valores do preditor seja aproximadamente uniforme dentro da amplitude de amostragem.* Cuidado com conjuntos de dados em que um ou dois dos valores da variável preditora sejam muito diferentes em grandeza dos outros. Tais pontos de influência podem dominar a inclinação da regressão e gerar relações significativas onde ela não existe de verdade (Figura 7.3; *ver* Capítulo 8 para discussão adicional sobre dados discrepantes). Algumas vezes, os pontos de influência podem

[2] O esquema de amostragem precisa refletir os objetivos do estudo. Se este for delineado simplesmente para documentar a relação entre a densidade de sementes e de roedores, então uma série de parcelas aleatórias podem ser selecionadas; a **correlação** é usada para explorar a relação entre as duas variáveis. No entanto, se a hipótese for de que a densidade de sementes seja responsável pela densidade de roedores, então uma série de parcelas que englobem uma gama uniforme de densidade de sementes deve ser amostrada; a **regressão** é usada para explorar a dependência funcional da abundância de roedores pela densidade de sementes. Idealmente, as parcelas amostradas devem diferir uma da outra apenas na densidade de sementes presente. Outra distinção importante é que uma verdadeira análise de regressão assume que o valor da variável independente seja exatamente conhecido e que não está sujeito a erros de medição. Por fim, a regressão linear padrão minimiza os desvios residuais apenas na direção vertical (y), enquanto correlação minimiza a distância perpendicular (x e y) de cada ponto até a linha de regressão (também chamada de regressão Modelo II). A distinção entre correlação e regressão é sutil e frequentemente confusa, porque algumas estatísticas (como o coeficiente de correlação) são idênticas para ambos os tipos de análises. *Ver* Capítulo 9 para mais detalhes.



Número da parcela	Sementes/m²	Cobertura de vegetação (%)	Roedores/m²
1	12.000	11	5,0
2	1.500	2	1,1
.	.	.	.
.	.	.	.
20	11.500	52	3,7

Figura 7.1 Arranjo espacial das réplicas para um estudo de regressão. Cada quadrado representa uma parcela de 25m² diferente. As parcelas foram amostradas para garantir uma cobertura uniforme da densidade de sementes (*ver* Figuras 7.2 e 7.3). Dentro de cada parcela, o investigador mede a densidade de roedores (a variável resposta) e a densidade de sementes e cobertura da vegetação (as duas variáveis preditoras). Os dados são organizados em uma planilha, na qual cada linha é uma parcela e as colunas são as variáveis medidas dentro da parcela.

ser corrigidos com uma transformação da variável preditora (*ver* Capítulo 8), mas ressaltamos que as análises não podem salvar um delineamento amostral ruim.

REGRESSÃO MÚLTIPLA A extensão para regressão múltipla é simples. Duas ou mais variáveis preditoras contínuas são medidas em cada réplica, junto com uma única variável resposta. Retornando ao exemplo dos roedores do deserto, você suspeita que, além da disponibilidade de sementes, a densidade de roedores também seja controlada pela estrutura da vegetação – em parcelas com

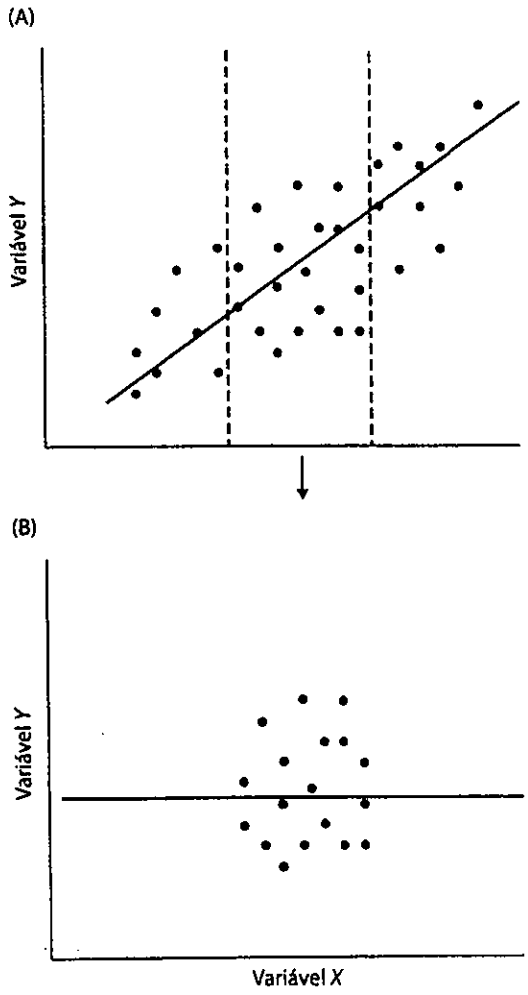


Figura 7.2 A amostragem inadequada sobre uma amplitude restrita da variável *X* pode criar uma inclinação de regressão falsamente não significativa, mesmo que o *X* e o *Y* sejam fortemente correlacionadas uma com a outra. Cada ponto representa uma réplica na qual um valor foi medido para as variáveis *X* e *Y*. Círculos verdes representam possíveis dados que não foram coletados para a análise. Círculos pretos representam as amostras das réplicas que foram medidas. (A) A amplitude total dos dados. A linha sólida indica a relação linear verdadeira entre as variáveis. (B) A linha de regressão é ajustada aos dados amostrados. Como a variável *X* foi amostrada em uma amplitude restrita de valores, a alteração na variável *Y* resultante é limitada, e a inclinação da regressão ajustada aparenta ser próxima de zero. Amostrar ao longo da amplitude total da variável *X* irá prevenir esse tipo de erro.

vegetação esparsa, os roedores do deserto são vulneráveis a aves predadoras (Abramsky et al., 1997). Neste caso, você poderia tomar três medidas em cada parcela: densidade de roedores, densidade de sementes e cobertura de vegetação. A primeira ainda é a variável resposta, as demais são as duas variáveis preditoras

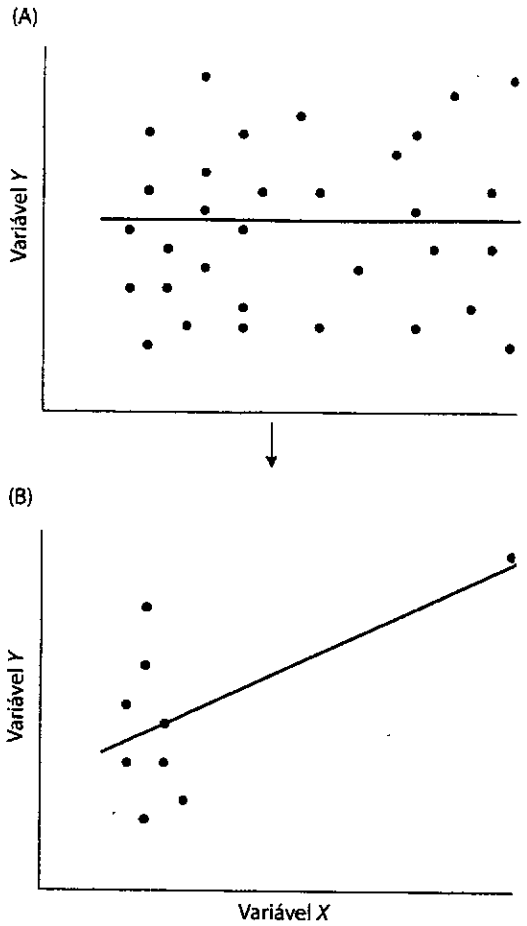


Figura 7.3 Falhas em amostrar uniformemente a amplitude total da variável podem levar a resultados falsos. Como na Figura 7.2, cada ponto preto representa uma observação registrada; os pontos verdes representam pares *X, Y* não observados. (A) A linha sólida indica a relação linear verdadeira entre as variáveis. Essa relação teria sido revelada se a variável *X* tivesse sido amostrada uniformemente. (B) A linha de regressão é ajustada só aos dados amostrados (i. e., apenas os pontos pretos). Como apenas um único dado com um valor alto de *X* foi medido, esse ponto teve uma influência excessiva sobre a linha de regressão ajustada. O resultado sugere, de forma inacurada, uma relação positiva entre as duas variáveis.

(Figura 7.1). Idealmente, as diferentes variáveis preditoras devem ser independentes uma da outra. Como em delineamentos de regressão simples, os valores diferentes das variáveis preditoras devem ser estabelecidos de maneira uniforme ao longo de toda a amplitude de valores possíveis. Isto é simples em um estudo experimental, mas raras vezes é possível em um estudo observacional. Nesse caso, é frequente a situação em que as variáveis preditoras serão correlacionadas entre si. Por exemplo, parcelas com alta densidade de vegetação provavelmen-

te terão alta densidade de sementes. Pode haver poucas, ou nenhuma, parcelas onde a densidade da vegetação seja alta e a de sementes, baixa (ou vice-versa). Essa **co-linearidade** torna difícil estimar, com acurácia, os parâmetros da regressão[3] e separar quanto da variação na variável resposta realmente está associada com cada uma das variáveis preditoras.

Como sempre, a replicação se torna importante ao passo que adicionamos mais variáveis preditoras na análise. Seguindo a Regra dos 10 (*ver* Capítulo 6), você deve tentar obter pelo menos 10 réplicas para cada variável preditora em seu estudo. Entretanto, em muitos casos, é muito mais fácil medir variáveis preditoras adicionais que obter réplicas independentes adicionais. No entanto, você deve resistir à tentação de medir tudo simplesmente porque é possível. Tente selecionar variáveis que são biologicamente importantes e relevantes para a hipótese ou a questão que você está tratando. É um erro pensar que um algoritmo de seleção de modelos, como uma regressão múltipla gradativa, pode identificar com segurança o conjunto de variáveis preditoras "correto" a partir de um grande conjunto de dados (Burnham & Anderson, 2002). Além disso, grandes conjuntos de dados frequentemente sofrem de **multicolinearidade**: muitas das variáveis preditoras são correlacionadas umas às outras (Graham, 2003).

Delineamentos de ANOVA

Se a sua variável preditora é categórica (ordenada ou não ordenada) e sua variável resposta é contínua, seu delineamento é chamado de **ANOVA** (do inglês **AN**alysis **O**f **VA**riance). A ANOVA também se refere às análises estatísticas desses tipos de delineamentos (*ver* Capítulo 10).

TERMINOLOGIA A ANOVA é repleta de terminologias. **Tratamentos** se referem às diferentes categorias das variáveis preditoras usadas. Em um estudo experimental, os tratamentos representam as diferentes manipulações realizadas. Em um estudo observacional, os tratamentos representam os diferentes grupos sob comparação. O número de tratamentos em um estudo é igual ao de categorias sob comparação. Dentro de cada tratamento, observações múltiplas serão feitas e cada uma é chamada de **réplica**. Em delineamentos de ANOVA padrão, cada réplica deve ser independente, tanto estatística quanto biologicamente, das outras réplicas dentro e entre os tratamentos. Mais adiante, neste capítulo, discutiremos certos delineamentos de ANOVA que atenuam o pressuposto de independência entre réplicas.

[3] De fato, se uma das variáveis preditoras pode ser descrita como uma função linear perfeita da outra, sequer será possível encontrar algebricamente os coeficientes da regressão. Mesmo quando o problema não é tão grave, as correlações entre variáveis preditoras tornam difícil o teste e a comparação de modelos. *Ver* MacNally (2000b) para uma discussão de variáveis correlacionadas e construção de modelos em biologia da conservação.

Também fazemos uma distinção entre **delineamentos com um fator** e **delineamentos multifatoriais**. No primeiro caso, cada um dos tratamentos representa a variação em uma única variável preditora ou fator. Cada valor do fator que representa um tratamento em particular é chamado de **nível do tratamento**. Por exemplo, um delineamento de ANOVA com um fator poderia ser utilizado para comparar as respostas do crescimento de plantas cultivadas em 4 níveis diferentes de nitrogênio, ou as respostas de crescimento de 5 espécies de plantas diferentes a um único nível de nitrogênio. Os grupos de tratamentos podem ser ordenados (p. ex., 4 níveis de nitrogênio) ou não ordenados (p. ex., 5 espécies de plantas).

No segundo tipo de delineamento, os tratamentos envolvem dois (ou mais) fatores diferentes e cada um é aplicado em combinação a diversos tratamentos. Em um delineamento multifatorial, existem diferentes níveis de tratamento para cada fator. Como no primeiro, neste caso, os tratamentos dentro de cada fator também podem ser ordenados ou não ordenados. Por exemplo, um delineamento de ANOVA de dois fatores poderia ser necessário se você quisesse comparar as respostas de plantas a 4 níveis de nitrogênio (Fator 1) *e* a 4 níveis de fósforo (Fator 2). Nesse delineamento cada um dos $4 \times 4 = 16$ níveis dos tratamentos representa uma combinação diferente de nível de nitrogênio e de fósforo. Cada combinação de nutrientes é aplicada a todas as réplicas dentro do tratamento (Figura 7.4).

Retornaremos a esse tópico mais adiante, porém vale a pena perguntar, aqui, qual a vantagem de usar um delineamento de dois fatores. Por que simplesmente não fazer dois experimentos separados? Por exemplo, você poderia testar os efeitos de fósforo em um delineamento de ANOVA de um fator com 4 níveis de tratamento e testar os efeitos de nitrogênio em um delineamento separado de ANOVA de um fator também com 4 níveis de tratamento. Qual a vantagem de usar um delineamento de dois fatores com 16 combinações de tratamentos de fósforo-nitrogênio em um único experimento?

Uma vantagem do delineamento de dois fatores é a eficiência. É provável que tenha maior custo-benefício fazendo apenas um experimento – mesmo com 16 tratamentos – do que ao fazer dois experimentos separados com 4 tratamentos cada. Uma vantagem mais importante é que delineamentos de dois fatores permitem testar tanto os efeitos principais (p. ex., os efeitos de nitrogênio e do fósforo sobre o crescimento das plantas) quanto os de interação (p. ex., interações entre nitrogênio e fósforo).

Os **efeitos principais** são os aditivos de cada nível de um tratamento, usando a média dos níveis do outro tratamento. Por exemplo, o efeito aditivo de nitrogênio deve representar a resposta das plantas em cada nível de nitrogênio, pela média das respostas aos níveis de fósforo. Ao contrário, o efeito aditivo do fósforo poderia ser medido como a resposta das plantas em cada nível de fósforo, pela média das respostas aos diferentes níveis de nitrogênio.

Os efeitos de interação representam respostas únicas a combinações de tratamentos específicas que não podem ser previstas simplesmente conhecendo os

Tratamento de nitrogênio (*esquema com um fator*)			
0,00 mg	0,10 mg	0,50 mg	1,00 mg
10	10	10	10

Tratamento de fósforo (*esquema com um fator*)			
0,00 mg	0,05 mg	0,10 mg	0,25 mg
10	10	10	10

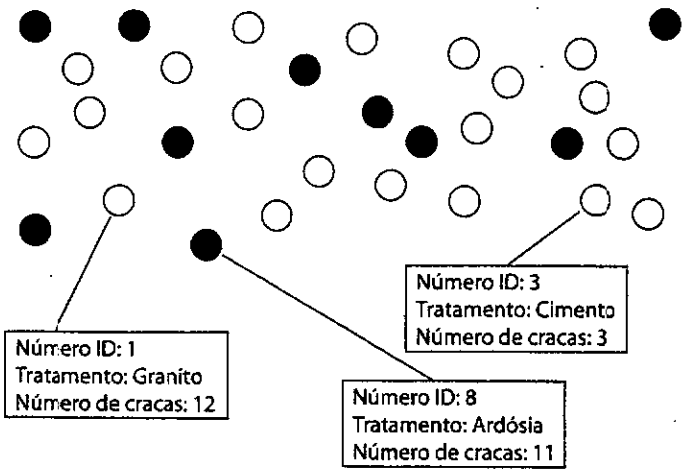
(Tratamentos simultâneos de N e P em um esquema de dois fatores)		Tratamento de nitrogênio			
		0,00 mg	0,10 mg	0,50 mg	1,00 mg
Tratamento de fósforo	0,00 mg	10	10	10	10
	0,05 mg	10	10	10	10
	0,10 mg	10	10	10	10
	0,25 mg	10	10	10	10

Figura 7.4 Combinações de tratamentos em delineamentos de um fator (os dois painéis superiores) e em um delineamento de dois fatores (painel inferior). Em todos os delineamentos, o número em cada célula indica o número de réplicas independentes a serem realizadas. Nos dois delineamentos de um fator (esquemas de uma via), os quatro níveis do tratamento representam quatro concentrações diferentes de nitrogênio e fósforo (mg/L). O tamanho amostral total é de 40 parcelas em cada experimento com um fator. No delineamento de dois fatores, os $4 \times 4 = 16$ tratamentos representam diferentes combinações de concentrações de nitrogênio e fósforo que são aplicadas de maneira simultânea em uma réplica. Esse delineamento de ANOVA completamente cruzado com 10 réplicas por combinação de tratamento requer um tamanho amostral total de 160 parcelas. *Ver* Figuras 7.9 e 7.10 para outros exemplos de delineamentos de dois fatores cruzados.

efeitos principais. Por exemplo, o crescimento de plantas em tratamento com muito nitrogênio e muito fósforo poderia ser sinergicamente maior do que você poderia prever conhecendo apenas os efeitos aditivos do nitrogênio e do fósforo em altos níveis. Os efeitos de interação frequentemente são a razão mais importante para usar um delineamento fatorial. Fortes interações são a força-guia por trás de muitas mudanças ecológicas e evolutivas, e com frequência são mais importantes que os efeitos principais. O Capítulo 10 discutirá métodos analíticos e termos de interação em mais detalhes.

ANOVA DE UM FATOR A ANOVA de um fator é um dos mais simples, porém robustos, delineamentos experimentais. Depois de descrever a estrutura de um fator básico, também explicaremos os delineamentos de blocos aleatorizados e de ANOVA aninhada. Estritamente falando, blocos aleatorizados e de ANOVA aninhada são delineamentos de dois fatores, mas o segundo fator (blocos, ou subamostras) é incluído apenas para controlar a variação amostral e não é de interesse primário.

O **esquema de um fator** é usado para comparar médias entre dois ou mais tratamentos ou grupos. Por exemplo, suponha que você queira determinar se o



Número ID	Tratamento	Réplica	Número de cracas
1	Granito	1	12
2	Ardósia	1	10
3	Cimento	1	3
4	Granito	2	14
5	Ardósia	2	10
6	Cimento	2	8
7	Granito	3	11
8	Ardósia	3	11
9	Cimento	3	7
.	.	.	.
.	.	.	.
30	Cimento	10	8

Figura 7.5 Exemplo de esquema de um fator. Tal experimento é delineado para testar o efeito do tipo de substrato sobre o recrutamento de cracas na zona intertidal rochosa (Caffey, 1982). Cada círculo representa um substrato rochoso independente. Existem 10 réplicas aleatoriamente distribuídas de cada um de três tratamentos, representados pelos três tons de verde. O número de cracas recrutadas é amostrado em um quadrado de 10 cm no centro de cada superfície rochosa. Os dados são organizados em uma planilha, na qual cada linha é uma réplica independente. As colunas indicam o número ID de cada réplica (1-30), o grupo de tratamento (cimento, ardósia ou granito), o número da réplica dentro de cada tratamento (1-10) e o número de cracas (a variável resposta).

recrutamento de cracas na zona intertidal em uma linha costeira é afetado pelos diferentes tipos de substratos rochosos (Caffey, 1982). Você pode começar obtendo um conjunto de ladrilhos de ardósia, granito e cimento. Os ladrilhos devem ser idênticos em tamanho e forma, e diferir apenas quanto ao material. Seguindo a Regra dos 10 (Capítulo 6), estabeleça 10 réplicas de cada tipo de substrato ($N = 30$ total). Cada réplica é colocada na zona intertidal intermediária, em um conjunto de coordenadas espaciais que foram escolhidas com um gerador de números aleatórios (Figura 7.5).

Depois de montar o experimento, você retorna após 10 dias e conta o número de cracas dentro de um quadrado de 10 cm × 10 cm centralizado em cada ladrilho. Os dados são organizados em uma planilha, na qual cada linha corresponde a uma réplica. As primeiras colunas contêm informações de identificação associadas com cada réplica, e a última coluna da planilha fornece o número de cracas no quadrado. Embora os detalhes sejam diferentes, esse é o mesmo esquema usado no estudo de densidade de formigas (descrito no Capítulo 5): muitas réplicas de observações independentes são obtidas para cada tratamento ou grupo amostral.

O esquema de um fator é um dos mais simples, embora dos mais poderosos, delineamentos experimentais, podendo ser facilmente adaptado para estudos em que o número de réplicas por tratamento não seja idêntico (tamanhos amostrais desiguais). O esquema de um fator permite-nos testar as diferenças entre os tratamentos, bem como hipóteses mais específicas acerca de quais médias de tratamentos em particular são diferentes ou similares (*ver* "Comparando Médias" no Capítulo 10).

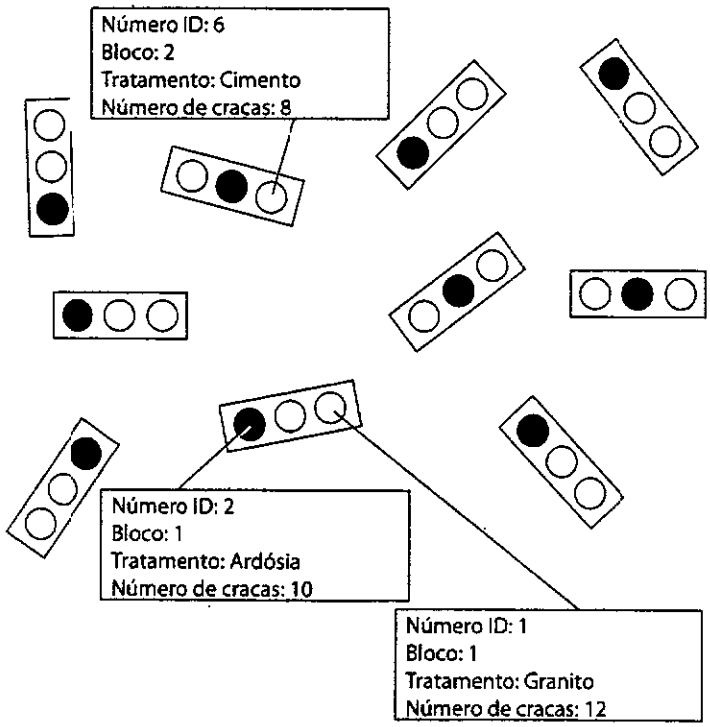
A principal desvantagem do esquema de um fator é que ele não acomoda explicitamente a heterogeneidade ambiental. A aleatorização completa das réplicas dentro de cada tratamento sugere que elas amostrarão todo o conjunto de condições possíveis, dentre as quais podem afetar a variável resposta. Por um lado, isso é positivo, pois significa que os resultados do experimento podem ser generalizados para todos esses ambientes. Por outro lado, se o "ruído" ambiental for muito mais forte que o "sinal" do tratamento, o experimento terá um baixo poder; a análise pode não revelar diferenças entre tratamentos, a menos que existam muitas réplicas. Outros delineamentos, incluindo o esquema de blocos aleatorizados e o de dois fatores, podem ser usados para acomodar a variabilidade ambiental.

Uma segunda, e mais sutil, desvantagem do esquema de um fator é que ele organiza os grupos de tratamentos ao longo de um único fator. Se os tratamentos representam diferentes tipos de fatores, então um esquema de dois fatores deve ser usado para separar os efeitos principais e os termos de interação. Termos de interação são especialmente importantes, porque o efeito de um fator com frequência depende dos níveis de outro. Por exemplo, o padrão de recrutamento em diferentes substratos pode depender dos níveis de um segundo fator (como a densidade de predadores).

DELINEAMENTOS DE BLOCOS ALEATORIZADOS Uma forma eficiente de incorporar a heterogeneidade ambiental é modificar a ANOVA de um fator e usar um **delineamento de blocos aleatorizados**. Um **bloco** é uma área delimitada ou um período de tempo dentro do qual as condições ambientais são relativamente homogêneas. Os blocos podem ser distribuídos aleatória ou sistematicamente na área de estudo, mas eles devem ser arranjados de forma que as condições ambientais sejam mais similares dentro dos blocos que entre eles.

Uma vez que os blocos estejam estabelecidos, as réplicas ainda terão que ser atribuídas de maneira aleatória aos tratamentos, mas há uma restrição nisso: que

uma única réplica de cada um dos tratamentos seja atribuída a cada bloco. Assim, em um delineamento de blocos aleatorizados simples, cada um contém exatamente uma réplica de todos os tratamentos do experimento. Dentro de cada bloco, a localização das réplicas dos tratamentos deve ser aleatorizada. A Figura 7.6 ilustra



Número ID	Tratamento	Bloco	Número de cracas
1	Granito	1	12
2	Ardósia	1	10
3	Cimento	1	3
4	Granito	2	14
5	Ardósia	2	10
6	Cimento	2	8
7	Granito	3	11
8	Ardósia	3	11
9	Cimento	3	7
.	.	.	.
.	.	.	.
30	Cimento	10	8

Figura 7.6 Exemplo de um delineamento de blocos aleatorizados. As 10 réplicas de cada um dos três tratamentos são agrupadas em blocos – agrupamentos físicos com uma réplica de cada um dos três tratamentos. Tanto a posição dos blocos quanto a dos tratamentos dentro dos blocos são aleatorizadas. A organização dos dados na planilha é idêntica à do esquema de um fator (Figura 7.5), mas a coluna de réplicas é substituída por outra, indicando o bloco ao qual cada réplica está associada.

o experimento das cracas como um delineamento de blocos aleatorizados. Como há 10 réplicas e 10 blocos (menos, se você replicar dentre de cada bloco) cada um conterá uma réplica de cada um dos três tratamentos. O esquema da planilha para esses dados é o mesmo que para o esquema de um fator, exceto com relação à coluna de réplicas, que agora é substituída por outra indicando o bloco.

Cada bloco deve ser pequeno o suficiente para abranger um conjunto de condições relativamente homogêneas. No entanto, cada um deve ser grande o suficiente para acomodar uma réplica de cada tratamento. Além disso, deve haver espaço dentro do bloco para um espaçamento entre as réplicas e a garantia de sua independência (*ver* Figura 6.5). Os próprios blocos também devem estar distantes um do outro o suficiente para garantir a independência das réplicas entre eles.

Se existirem gradientes geográficos nas condições ambientais, então cada bloco deve abranger um pequeno intervalo do gradiente. Por exemplo, existem fortes gradientes ambientais ao longo da encosta de uma montanha. Desse modo, podemos estabelecer um experimento com três blocos, cada um em uma altitude – baixa, média e alta (Figura 7.7A). Porém, não seria apropriado criar três blocos no mesmo sentido do gradiente – da maior altitude para a mais baixa (Figura 7.7B)–, cada bloco englobaria condições muito heterogêneas. Em outros casos, a variação ambiental pode ocorrer em manchas, e os blocos devem ser arranjados de forma a refleti-las. Por exemplo, se um experimento for conduzido em um complexo de áreas alagadas, cada charco semi-isolado poderia ser tratado como um bloco. Por fim, se a organização espacial da heterogeneidade ambiental for desconhecida, os blocos podem ser arranjados de maneira aleatória dentro da área de estudo.[4]

O delineamento de blocos aleatorizados é um modo eficiente e bastante flexível que fornece um controle simples para a heterogeneidade ambiental. Ele pode ser usado para controlar gradientes ambientais e manchas de hábitat. Como veremos no Capítulo 10, quando houver heterogeneidade ambiental presente, o delineamento de blocos aleatorizados é mais eficiente que um esquema de um fator completamente aleatorizado, que pode requerer muito mais replicação para obter o mesmo poder estatístico.

O delineamento de blocos aleatorizados é também útil quando sua replicação é restringida por espaço ou tempo. Por exemplo, suponha que você

[4] O delineamento de blocos aleatorizados permite que você estabeleça seus blocos para abranger gradientes ambientais em uma única dimensão espacial. Mas e se a variação ocorre em duas dimensões? Por exemplo, suponha que existe um gradiente de umidade do Norte para o Sul em um campo, mas também exista um gradiente do Leste para o Oeste na densidade de predadores. Em tais casos, delineamentos de blocos aleatorizados mais complexos podem ser usados. Por exemplo, o **quadrado latino** é um delineamento de blocos em que os n tratamentos são colocados no campo em um quadrado $n \times n$; cada tratamento aparece exatamente uma vez em todas as linhas e uma vez em todas as colunas do esquema. Sir Ronald Fisher (*ver* Nota de Rodapé 5, Capítulo 5) foi pioneiro nesses tipos de delineamentos para estudos de agronomia, nos quais um único campo é dividido e os tratamentos aplicados às subparcelas contíguas. Esses delineamentos não têm sido utilizados com frequência por ecólogos, por que restrições de aleatorizações e de esquematização são difíceis de obter em experimentos de campo.

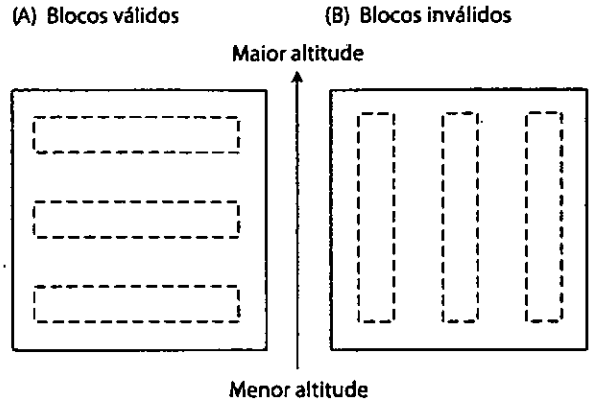


Figura 7.7 Delineamentos de bloco válidos e inválidos. (A) Três blocos devidamente orientados, cada um englobando a mesma elevação em uma encosta de montanha ou outro gradiente ambiental qualquer. As condições ambientais são mais similares dentro do que entre blocos. (B) Esses blocos são orientados indevidamente, indo "no sentido" do gradiente de altitude. Condições são tão heterogêneas dentro dos blocos quanto entre eles, então nenhuma vantagem é obtida em seu uso.

está fazendo um experimento de laboratório sobre crescimento de algas com 8 tratamentos e quer completar 10 réplicas por tratamento. No entanto, você tem espaço suficiente em seu laboratório para fazer apenas 12 réplicas por vez. Como proceder? Você deve fazer o experimento em "blocos", montando uma única réplica de cada um dos 8 tratamentos. Após registrar os resultados, monte o experimento de novo (incluindo outro conjunto de aleatorizações para o estabelecimento e posicionamento dos tratamentos) e continue até acumular 10 blocos. Esse delineamento controla as mudanças inevitáveis nas condições ambientais que ocorrem em seu laboratório, ao longo do tempo, e ainda permite comparações apropriadas dos tratamentos. Em outros casos, a limitação pode não ser de espaço, mas de organismos. Por exemplo, em um estudo sobre o acasalamento de peixes, você pode ter que esperar até que tenha um certo número de peixes sexualmente maduros antes de montar e executar um único bloco do experimento. Em ambos os exemplos, o delineamento de blocos aleatorizados é a melhor precaução contra a variação nas condições durante o decorrer de seu experimento.

Por fim, o delineamento de blocos aleatorizados pode ser adaptado para um **esquema de pares casados**. Cada bloco consiste em um grupo de organismos individuais ou de parcelas que foram deliberadamente escolhidas para terem as características mais similares possíveis. Cada réplica no grupo recebe a atribuição de um dos tratamentos. Por exemplo, em um simples estudo experimental dos efeitos da abrasão sobre o crescimento de corais, um par de cabeças de corais com tamanhos similares poderia ser considerado um único bloco. Uma das cabeças de coral poderia ser aleatoriamente atribuída ao grupo controle e a outra, ao grupo de abrasão. Outro par casado poderia ser escolhido da mesma forma

e os tratamentos aplicados. Mesmo os indivíduos em cada par não sendo parte de um bloco espacial ou temporal, eles provavelmente serão mais similares que indivíduos em qualquer outro bloco, pois foram casados com base no tamanho da colônia ou em outras características. Por essa razão, a análise utilizará um delineamento de blocos aleatorizados. A abordagem de pares casados é um método muito eficiente quando as respostas das réplicas são muito heterogêneas. Parear os indivíduos controla essa heterogeneidade, tornando mais fácil detectar os efeitos do tratamento.

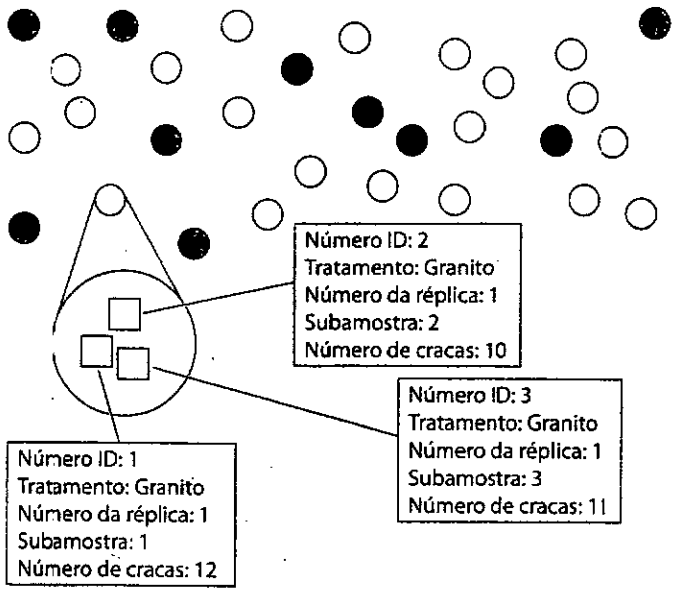
Existem quatro desvantagens para o delineamento de blocos aleatorizados. Na primeira, existe um custo estatístico para fazer o experimento com blocos. Se o tamanho amostral for pequeno e o efeito do bloco fraco, o delineamento de blocos aleatorizados é menos robusto que um esquema de um fator (*ver* Capítulo 10). A segunda desvantagem é que, se blocos forem muito pequenos, você pode introduzir falta de independência ao aglomerar fisicamente os tratamentos. Como discutimos no Capítulo 6, aleatorizar a posição dos tratamentos dentro do bloco ajudará com tal problema, mas não o eliminará inteiramente. A terceira desvantagem do delineamento de blocos aleatorizados é que, se uma das réplicas for perdida, os dados daquele bloco não podem ser usados, a menos que os valores não disponíveis possam ser estimados de maneira indireta.

A quarta – e mais séria – desvantagem do delineamento de blocos aleatorizados é que ele assume que não há interação entre blocos e tratamentos. O delineamento de blocos leva em conta as diferenças aditivas na variável resposta e assume que a ordem do ranque das respostas ao tratamento não muda de um bloco para o próximo. Voltando para o exemplo das cracas, o modelo de blocos aleatorizados assume que se o recrutamento em um dos blocos for alto, todas as observações naquele bloco serão assim. No entanto, assume-se que os efeitos do tratamento sejam consistentes de um bloco para o próximo, de forma que a ordem do ranque no recrutamento de cracas entre tratamentos (granito > ardósia > cimento) seja a mesma, independente de quaisquer diferenças nos níveis gerais de recrutamento entre blocos. Mas suponhamos que em alguns blocos o recrutamento seja maior nos substratos de cimento e em outros, nos de granito. Neste caso, o delineamento de blocos aleatorizados pode falhar em caracterizar de maneira adequada os efeitos principais do tratamento. Por essa razão, alguns autores (Mead, 1988, Underwood, 1997) têm argumentado que o simples delineamento de blocos aleatorizados não deve ser usado, a menos que exista replicação dentro dos blocos. Assim, o delineamento se torna uma análise de variância de dois fatores, que discutiremos a seguir.

A replicação dentro de blocos de fato irá separar os efeitos principais, os efeitos dos blocos e a interação entre blocos e tratamentos. A replicação também irá tratar o problema de dados faltantes ou perdidos dentro de um bloco. No entanto, os ecólogos frequentemente não têm o luxo de replicação dentro de blocos, em especial quando o fator do bloco não é de interesse primário. O delineamento de blocos aleatorizados simples (sem replicação) irá, no mínimo, captar o componente aditivo (em geral o componente mais importante) de va-

riação ambiental que poderia, de outro modo, ser aglomerado ao erro puro em um esquema de um fator.

DELINEAMENTOS ANINHADOS **Um delineamento aninhado** se refere a qualquer delineamento em que há subamostragem dentro de cada uma das réplicas. Conforme o exemplo das cracas, suponha que, em vez de medir o recrutamento para uma réplica em um único quadrado de 10 × 10 cm, você decidiu fazer três dessas medidas em cada um dos 30 ladrilhos do estudo (Figura 7.8). Embora o



Número ID	Tratamento	Número da réplica	Subamostra	Número de cracas
1	Granito	1	1	12
2	Granito	1	2	10
3	Granito	1	3	11
4	Ardósia	2	1	14
5	Ardósia	2	2	10
6	Ardósia	2	3	7
7	Cimento	3	1	5
8	Cimento	3	2	6
9	Cimento	3	3	10
.	.	.	.	.
.	.	.	.	.
90	Cimento	30	3	6

Figura 7.8 Exemplo de um delineamento aninhado. O estudo é o mesmo mostrado nas Figuras 7.5 e 7.6. O esquema é idêntico ao de um fator da Figura 7.5, mas neste caso três subamostras são realizadas em cada réplica independente. Na planilha, uma coluna extra é adicionada para indicar o número da subamostra, e o total de observações aumentou de 30 para 90.

número de réplicas não tenha aumentado, o de observações aumentou de 30 para 90. Agora, cada linha na planilha desses dados representa uma subamostra diferente e as colunas indicam de qual réplica e de qual tratamento a subamostra foi retirada.

Este é o primeiro delineamento no qual incluímos subamostras que claramente não são independentes umas das outras. Qual é a lógica para tal esquema amostral? A razão principal é aumentar a precisão com que estimamos a resposta para cada réplica. Por causa da Lei dos Grandes Números (Capítulo 3), quanto mais subamostras usarmos, mais precisamente estimaremos a média para cada réplica. O aumento na precisão deve aumentar o poder do teste.

Existem três vantagens em usar um delineamento aninhado. A primeira é que a subamostragem aumenta a precisão da estimativa para cada réplica no delineamento. Na segunda, o delineamento aninhado permite testar duas hipóteses: primeira, existe variação entre tratamentos? E, segunda, existe variação entre as réplicas *dentro* de um tratamento? A primeira hipótese é equivalente a um delineamento de um fator que usa as *médias das subamostras* como a observação para cada réplica. A segunda hipótese é equivalente ao delineamento de um fator que usa as subamostras para testar por diferenças entre réplicas *dentro* dos tratamentos.[5]

Por fim, o delineamento aninhado pode ser estendido para um tipo amostral hierárquico. Por exemplo, você poderia, em um único estudo, fazer contagens em subamostras aninhadas dentro de réplicas, réplicas aninhadas dentro de zonas intertidais, zonas intertidais aninhadas dentro de costões, costões aninhados dentro de regiões, e até mesmo regiões aninhadas dentro de continentes (Caffey, 1985). A razão para realizar esse tipo de amostragem é que podemos fazer a partição da variação nos dados em componentes que representam cada um dos níveis hierárquicos do estudo (*ver* Capítulo 10). Por exemplo, você pode ser capaz de mostrar que 80% da variação nos dados ocorrem ao nível de zonas intertidais dentro de costões, mas apenas 2% podem ser atribuídas à variação entre costões dentro de uma região. Isto pode significar que a densidade de cracas varia fortemente da zona intertidal alta para a baixa, mas não varia muito de uma linha costeira para outra. Tais afirmações são úteis para acessar a importância relativa de diferentes mecanismos em produzir padrões (Petraitis, 1998; *ver também*, Figura 4.6).

Os delineamentos aninhados são potencialmente perigosos, ao passo que eles com frequência são analisados de maneira incorreta. Um dos equívocos mais sérios e comuns em ANOVA é os investigadores tratarem cada subamostra como uma réplica independente e analisar o delineamento aninhado como o de um

[5]Você pode pensar nesta segunda hipótese como um delineamento de um fator em um nível hierárquico inferior. Por exemplo, suponha que você usou os dados apenas das quatro réplicas do tratamento de granito. Considere cada uma como um "tratamento" diferente e cada subamostra como uma "réplica" diferente para aquele tratamento. O delineamento agora é de um fator que compara as réplicas do tratamento de granito.

fator (Hurlbert, 1984). A falta de independência das subamostras aumenta artificialmente o tamanho amostral (o triplo em nosso exemplo, no qual tiramos três subamostras em cada ladrilho) e infla a probabilidade de um erro estatístico do Tipo I (i. e., erroneamente rejeitar uma hipótese nula verdadeira). Um segundo e menos sério problema é que o delineamento aninhado pode ser difícil ou até impossível de analisar de maneira adequada se o tamanho amostral não for o mesmo em cada grupo. Mesmo com um número idêntico de amostras e de subamostras, a amostragem aninhada em esquemas mais complexos, como o de dois fatores ou o delineamento parcelas subdivididas, pode ser complicado para analisar. As configurações padrão dos programas estatísticos em geral não são apropriadas.

Contudo, a desvantagem mais grave do delineamento aninhado é que ele frequentemente representa um caso de má alocação do esforço amostral. Como veremos no Capítulo 10, o poder dos delineamentos de ANOVA dependem muito mais do número de réplicas independentes do que da precisão com a qual cada réplica é medida. É uma estratégia muito melhor investir seu esforço amostral em obter mais réplicas independentes que subamostrar dentro de cada réplica. Especificando com cuidado seu protocolo de amostragem (p. ex., "apenas frutos não danificados de plantas não aglomeradas crescendo sob sombreamento completo"), você pode ser capaz de aumentar a precisão de suas estimativas com mais eficiência que com repetidas subamostragens.

Dito isto, você certamente deve continuar a subamostrar, caso seja rápido e barato. No entanto, nosso conselho é que você tire a média (ou junte) dessas subamostras, de forma que você tenha uma única observação para cada réplica e, em seguida, trate o experimento como um delineamento de um fator. Uma vez que os números não sejam muito oscilantes, tirar a média pode também aliviar problemas de tamanhos amostrais desiguais entre subamostras, além de melhorar o ajuste dos erros a uma distribuição normal. É possível, porém, que, após tirar a média entre subamostras dentro das réplicas, você não tenha mais réplicas o suficiente para uma análise completa. Neste caso, é necessário um delineamento com mais réplicas que sejam verdadeiramente independentes. Subamostrar não é solução para uma replicação inadequada!

DELINEAMENTOS MULTIFATORIAIS: ESQUEMA DE DOIS FATORES Os delineamentos multifatoriais estendem os princípios do esquema de um fator para dois ou mais fatores de tratamento. As questões sobre aleatorização, esquemas e amostragem são idênticas àquelas discutidas para os delineamentos de um fator de blocos aleatorizados e aninhados. De fato, a única diferença real no delineamento está na atribuição dos tratamentos a dois ou mais fatores em vez de um único fator. Como antes, os fatores podem representar tanto tratamentos ordenados quanto não ordenados.

Retornando ao exemplo das cracas, suponhamos que, além do efeito do substrato, você queria testar os efeitos de caramujos predadores sobre o recrutamento de cracas. Você poderia estabelecer um segundo experimento de um fator em

quatro tratamentos: sem manipulação, gaiola controle,[6] exclusão do predador e inclusão do predador. No entanto, em vez de rodar dois experimentos separados, você decide examinar ambos os fatores em um único experimento. Isso não é apenas um uso mais eficiente do seu tempo de campo, mas também o efeito dos predadores sobre o recrutamento de cracas pode ser diferente, dependendo do tipo de substrato. Portanto, você estabelece tratamentos em que aplica, de forma simultânea, um substrato e um tratamento de predação diferentes.

Este é um exemplo de **delineamento fatorial**, no qual dois ou mais fatores são testados simultaneamente em um experimento. O elemento-chave de um delineamento fatorial adequado é que os tratamentos são **completamente cruzados** e **ortogonais**: todo nível de tratamento do primeiro fator (substrato) deve ser representado com todos os níveis do segundo fator (predação; Figura 7.9). Assim, o experimento de dois fatores tem $3 \times 4 = 12$ combinações distintas de tratamentos, em oposição a apenas 3 tratamentos para um experimento de um fator de substrato ou 4 tratamentos para um experimento de um fator de predação. Note que cada um desses experimentos de um fator ficaria restrito a apenas uma das combinações de tratamentos para o outro fator. Em outras palavras, o experimento de substrato que descrevemos anteriormente foi conduzido sem manipulação para o tratamento de predação, o qual poderia ser conduzido em apenas um tipo de substrato. Uma vez que as combinações de tratamentos tenham sido determinadas, o estabelecimento físico do experimento seria o mesmo que para um esquema de um fator com 12 combinações de tratamentos (Figura 7.10).

No experimento de dois fatores, é crítico que todas as combinações cruzadas de tratamentos estejam representadas no delineamento. Se alguma das combinações de tratamentos estiver faltando, teremos um delineamento confundido. Como exemplo extremo, suponha que montamos apenas os tratamentos substrato de granito-exclusão de predadores e o de substrato de ardósia-inclusão de predadores. Agora, o efeito do predador está confundido com o do substrato. Se os resultados forem estatisticamente significantes ou não, não podemos separar quando o padrão é devido ao efeito do predador, ao efeito do substrato, ou à interação entre eles.

Tal exemplo ressalta uma diferença importante entre experimentos manipulativos e estudos observacionais. No estudo observacional, podemos coletar dados da variação na abundância de predadores e de presas a partir de uma gama

Gaiola e gaiola-controle

[6] Em uma gaiola-controle, os investigadores buscam simular as condições físicas geradas pela gaiola, mas ainda permitem que os organismos se movam livremente dentro e fora da parcela. Por exemplo, uma gaiola-controle pode consistir de um teto telado (colocada sobre uma parcela) que permita a entrada de caramujos predadores pelos lados. Em um tratamento de exclusão, todos os predadores são removidos de uma gaiola telada. No tratamento de inclusão, os predadores são colocados dentro de cada gaiola telada. A figura ao lado ilustra uma gaiola (painel superior) e uma gaiola-controle (painel inferior) em um experimento de exclusão de peixes em um riacho venezuelano (Flecker, 1996).

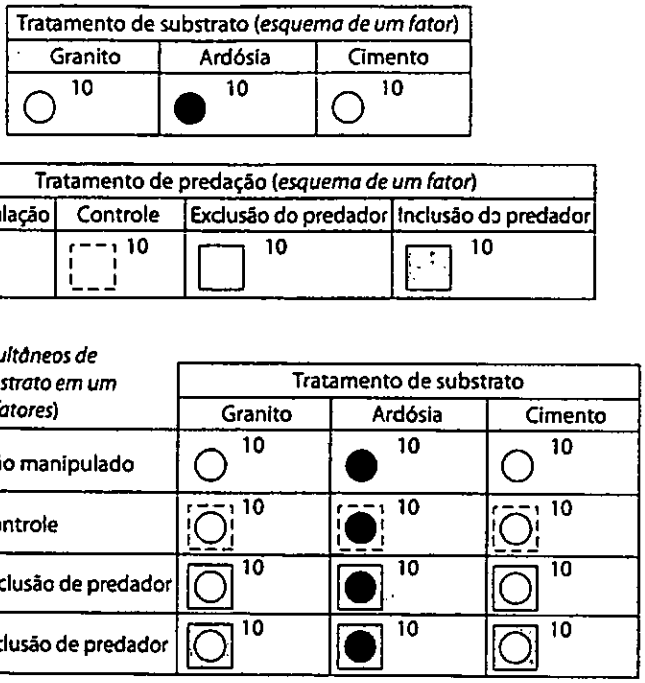
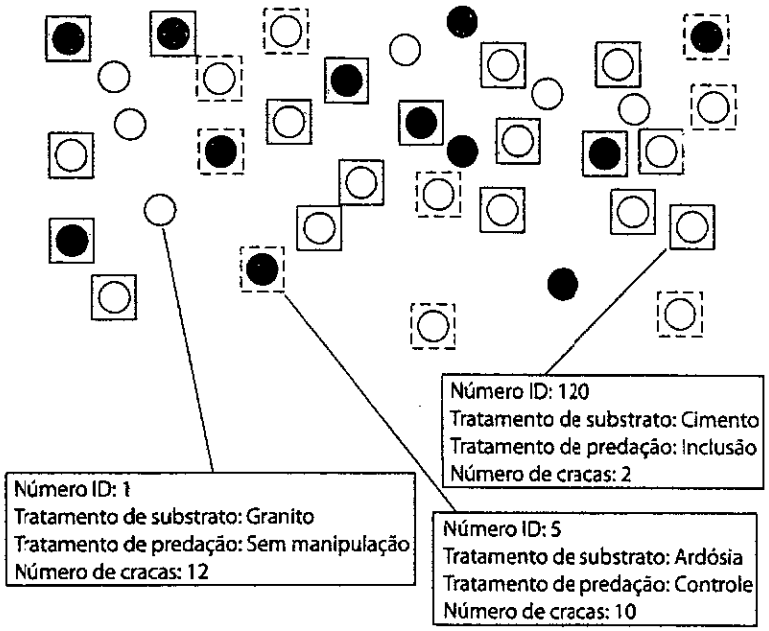


Figura 7.9 Combinações de tratamentos em dois delineamentos de um fator e em um de dois fatores completamente cruzados. Este experimento é delineado para testar o efeito dos tipos de substratos (granito, ardósia ou cimento) e de predação (sem manipulação, controle, exclusão do predador, inclusão do predador) sobre o recrutamento de cracas na zona intertidal rochosa. O número 10 indica o total de réplicas em cada tratamento. Os três tons dos círculos representam os três tratamentos de substratos e os padrões dos quadrados, os quatro tratamentos de predação. Os dois painéis superiores ilustram os dois delineamentos de um fator, em que apenas um dos dois fatores é variado sistematicamente. No de dois fatores (painel inferior), os $4 \times 3 = 12$ tratamentos representam diferentes combinações de substrato e predação. O símbolo em cada célula indica a combinação de tratamentos de predação e substrato que foram aplicados.

de amostras. Mas os predadores frequentemente estão restritos a apenas certos micro-hábitats ou tipos de substratos, de forma que a presença ou ausência do predador seja de fato naturalmente confundida com diferenças nos tipos de substratos. Isto torna difícil separar causa e efeito (*ver* Capítulo 6). O forte dos experimentos de campo multifatoriais é que eles desmembram essa co-variação natural e revelam os efeitos de múltiplos fatores separadamente e em conjunto. O fato de que algumas dessas combinações de tratamentos possam ser artificiais e raras vezes, se não nunca, encontradas na natureza, na verdade é um dos potenciais do experimento: ele revela a contribuição independente de cada fator aos padrões observados.

A principal vantagem dos delineamentos de dois fatores é a habilidade de separar efeitos principais e as interações entre esses fatores. Como discutiremos



Número ID	Tratamento de substrato	Tratamento de predação	Número de cracas
1	Granito	Sem manipulação	12
2	Ardósia	Sem manipulação	10
3	Cimento	Sem manipulação	8
4	Granito	Controle	14
5	Ardósia	Controle	10
6	Cimento	Controle	8
7	Granito	Exclusão de predador	50
8	Ardósia	Exclusão de predador	68
9	Cimento	Exclusão de predador	39
.	.	.	.
.	.	.	.
120	Cimento	Inclusão de predador	2

Figura 7.10 Exemplo de um delineamento de dois fatores. Os símbolos dos tratamentos são dados na Figura 7.9. A planilha contém colunas para indicar qual tratamento de substrato e qual de predação foram aplicados em cada réplica. O delineamento inteiro inclui $4 \times 3 \times 10 = 120$ réplicas no total, mas apenas 36 réplicas (três por combinação de tratamentos) foram ilustradas.

no Capítulo 10, o termo de interação representa o componente não aditivo da resposta. A interação mede a extensão na qual as diferentes combinações de tratamentos agem aditiva, sinérgica ou antagonicamente.

Talvez a principal desvantagem do delineamento de dois fatores seja que o número de combinações de tratamentos pode; com rapidez, tornar-se muito grande para uma replicação adequada. No exemplo de predação das cracas,

120 réplicas seriam necessárias para replicar 10 vezes cada combinação de tratamento.

Como no delineamento de um fator, um simples esquema de dois fatores não leva em conta a heterogeneidade espacial. Isso pode ser tratado por um delineamento de blocos aleatorizados simples, no qual cada bloco contém exatamente uma das combinações de tratamentos. Alternativamente, se você replicar todos os tratamentos dentro de cada bloco, isto se tornaria um delineamento de três fatores, com os blocos formando o terceiro fator na análise.

Uma limitação final dos delineamentos de dois fatores é que pode não ser possível estabelecer todas as combinações ortogonais de tratamentos. É um tanto surpreendente que, para vários experimentos ecológicos corriqueiros, o conjunto total de combinações de tratamentos pode não ser viável ou lógico. Por exemplo, suponha que você esteja estudando os efeitos da competição entre duas espécies de salamandras sobre a taxa de sobrevivência destes animais. Você decide usar um simples delineamento de dois fatores no qual cada espécie representa um dos fatores. Dentro de cada um, os dois tratamentos são a presença ou a ausência da espécie. Esse delineamento completamente cruzado produz quatro tratamentos (Tabela 7.2). Mas o que você medirá na combinação de tratamentos em que não tenha nem a espécie A nem a espécie B? Por definição, não há nada para medir neste caso. Em vez disso, você terá que estabelecer os outros três tratamentos ([Espécie A Presente, Espécie B Ausente], [Espécie A Ausente, Espécie B Presente], [Espécie A Presente, Espécie B Presente]) e analisar o delineamento como uma ANOVA de um fator. O delineamento de dois fatores seria possível apenas se mudarmos a variável resposta. Se a variável resposta for a abundância de presas comidas por salamandras, em vez da sobrevivência das salamandras, podemos então estabelecer o tratamento sem salamandras e medir os níveis de presa no esquema de dois

TABELA 7.2 Combinação de tratamentos em um esquema de dois fatores para experimentos com a simples adição ou remoção de espécies

	Espécie A	
Espécie B	**Ausente**	**Presente**
Ausente	10	10
Presente	10	10

O valor em cada célula é o número de réplicas de cada combinação de tratamentos. Se a variável resposta for alguma propriedade da própria espécie (p. ex., sobrevivência, taxa de crescimento), então a combinação Espécie A Ausente-Espécie B Ausente (na caixa) logicamente não é possível, e a análise terá que usar um esquema de um fator com três grupos de tratamentos (Espécie A Presente-Espécie B Presente, Espécie A Presente-Espécie B Ausente e Espécie A Ausente-Espécie B Presente). Se a variável resposta for alguma propriedade do ambiente que potencialmente seja afetada pelas espécies (p. ex., abundância de presas, pH), então as quatro combinações de tratamentos podem ser usadas e analisadas conforme uma ANOVA com dois fatores ortogonais (Espécie A e Espécie B), cada uma com dois níveis de tratamento (Ausente, Presente).

fatores completamente cruzados. Certamente, agora esse experimento faz uma pergunta bastante diferente.

Os experimentos de competição entre duas espécies, como nosso exemplo das salamandras, têm uma longa história na pesquisa ecológica e ambiental (Goldberg & Scheiner, 2001). Um número de problemas sutis surge no delineamento e na análise de experimentos de competição entre duas espécies. Esses experimentos tentam distinguir entre uma espécie focal, para a qual a variável resposta é medida, uma espécie associativa, cuja densidade é manipulada, e espécies não focais, que podem estar presentes, mas não são manipuladas experimentalmente.

A primeira questão é qual tipo de delineamento utilizar: **aditivo, substitutivo** ou **superfície de resposta** (Figura 7.11; Silvertown, 1987). Em um delineamento aditivo, a densidade da espécie focal é mantida constante enquanto a densidade da espécie experimental é variada. No entanto, esse delineamento confunde os efeitos de densidade e de frequência. Por exemplo, se compararmos uma parcela

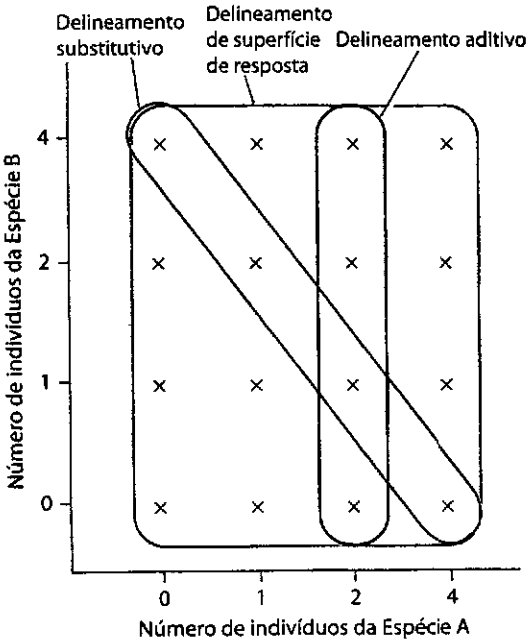


Figura 7.11 Delineamentos experimentais para experimentos de competição. A abundância de cada uma das Espécies A e B é estabelecida em 0, 1, 2, ou 4 indivíduos. Cada X indica uma combinação de tratamentos diferentes. Em um delineamento aditivo, a abundância de uma espécie é fixa (2 indivíduos da Espécie A) e a abundância do competidor varia (0, 1, 2, ou 4 indivíduos da Espécie B). Em um delineamento substitutivo, a abundância total de ambos competidores é mantida constante com 4 indivíduos, mas a composição de espécies nos diferentes tratamentos é alterada (0,4; 1,3; 2,2; 3,1; 4,0). Em um delineamento de superfície de resposta, todas as combinações de abundância dos dois competidores são estabelecidas em tratamentos diferentes ($4 \times 4 = 16$ tratamentos). O delineamento de superfície de resposta é o preferido, porque segue o princípio de uma boa ANOVA de dois fatores: os níveis do tratamento são completamente ortogonais (todos os níveis de abundância da Espécie A são representados com todos os da Espécie B).(*Ver* Inouye, 2001, para mais detalhes. Figura modificada a partir de Goldberg & Scheiner, 2001.)

controle (5 indivíduos da Espécie A, 0 indivíduos da Espécie B) com uma de adição (5 indivíduos da Espécie A, 5 indivíduos da Espécie B), teremos confundido a densidade total (10 indivíduos) com a presença do competidor (Underwood, 1986, Bernardo et al., 1995). Por outro lado, alguns autores têm argumentado que tais mudanças na densidade são de fato observadas quando uma nova espécie entra em uma comunidade e estabelece uma população, de forma que ajustar pela densidade total não necessariamente seja apropriado (Schluter, 1995).

Em um delineamento substitutivo, a densidade total dos organismos é mantida constante, mas as proporções relativas dos dois competidores são variáveis. Esses delineamentos medem a intensidade relativa da competição inter e intraespecífica, mas não medem a força absoluta da competição e assumem que as respostas são comparáveis em diferentes níveis de densidade.

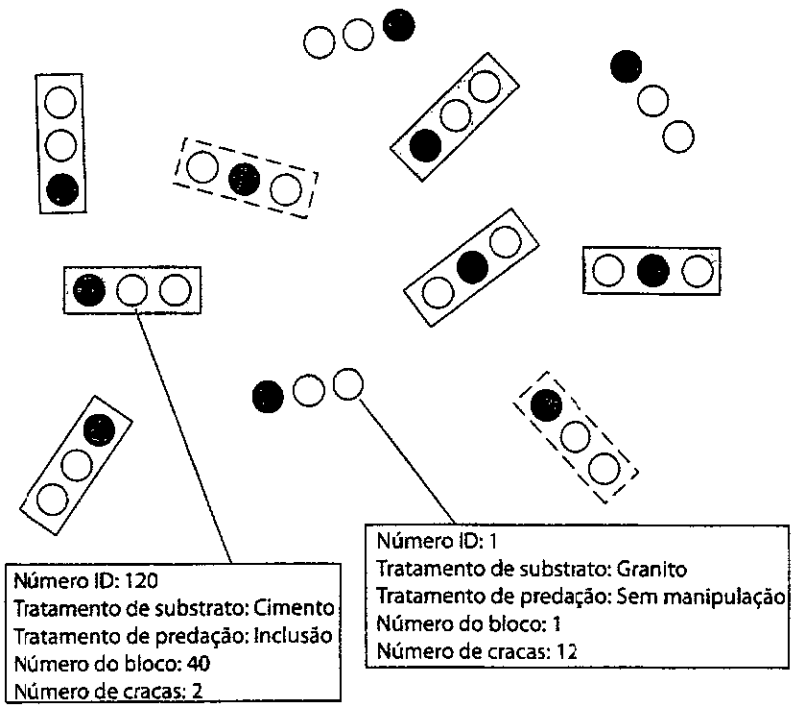
O delineamento de superfície de resposta correspode a um de dois fatores completamente cruzados, que varia tanto a proporção relativa quanto a densidade de competidores. Pode ser usado para medir a intensidade relativa e a força absoluta de interações competitivas inter e intraespecíficas. No entanto, como em todos experimentos de dois fatores com muitos níveis de tratamento, a replicação adequada pode ser um problema. Inouye (2001) revisa com minúcia os delineamentos de superfície de resposta e outras alternativas para estudos de competição.

Outras questões que precisam ser consideradas em experimentos de competição incluem: quantos níveis de densidade incorporar no intuito de estimar com acurácia efeitos competitivos; como lidar com a falta de independência de indivíduos dentro de uma réplica de tratamento; quando manipular ou controlar as espécies não focais; e como lidar com efeitos persistentes (*carry over*) e com a heterogeneidade espacial que são gerados em experimentos de remoção, nos quais as parcelas são estabelecidas com base na presença de uma espécie (Goldberg & Scheiner, 2001).

DELINEAMENTOS DE PARCELAS SUBDIVIDIDAS O delineamento de parcelas subdivididas é uma extensão daquele de blocos aleatorizados para dois tratamentos experimentais. A terminologia vem de estudos de agronomia, nos quais uma única parcela é dividida em subparcelas, cada uma recebendo um tratamento diferente. Para nosso objetivo, tal parcela subdividida é equivalente a um bloco composto internamente de réplicas dos diferentes tratamentos.

O que distingue um delineamento de parcelas subdivididas de um de blocos aleatorizados é que um segundo fator de tratamento também é aplicado – desta vez, no nível da parcela inteira. Voltemos ao exemplo das cracas. Mais uma vez, você vai estabelecer um delineamento de dois fatores, testando os efeitos da predação e do substrato. No entanto, suponha que as gaiolas sejam caras, consomem tempo para construir e você suspeite que exista muita variação de micro-hábitat no ambiente que esteja afetando seus resultados. Em um delineamento de parcelas subdivididas você poderia agrupar os três substratos, exatamente como fez no delineamento de blocos aleatorizados. No entanto, poderia, então, colocar uma *única gaiola* sobre todas as três réplicas de substratos dentro de um único bloco. Neste caso, o tratamento de predação é tratado como **fator de parcela completa,** porque um único tratamento de predação é aplicado a um bloco inteiro. O tratamento de substrato

é tratado como o **fator de subparcela**, porque todos os tratamentos de substrato são aplicados dentro de um único bloco. O delineamento de parcelas subdivididas é ilustrado na Figura 7.12.



Número ID	Tratamento de substrato	Tratamento de predação	Número do bloco	Número de cracas
1	Granito	Sem manipulação	1	12
2	Ardósia	Sem manipulação	1	10
3	Cimento	Sem manipulação	1	8
4	Granito	Controle	2	14
5	Ardósia	Controle	2	10
6	Cimento	Controle	2	8
7	Granito	Exclusão de predador	3	50
8	Ardósia	Exclusão de predador	3	68
9	Cimento	Exclusão de predador	3	39
.	.	.	.	.
.	.	.	.	.
120	Cimento	Inclusão de predador	40	2

Figura 7.12 Exemplo de um delineamento de parcelas subdivididas. Os símbolos dos tratamentos são dados na Figura 7.9. Os três tratamentos de substrato (fator de subparcela) são agrupados em blocos. O tratamento de predação (fator de parcela completa), em um bloco inteiro. A planilha contém colunas para indicar o tratamento de substrato, o de predação e a identificação do bloco para cada réplica. Apena um subconjunto dos blocos em cada tratamento de predação é ilustrado. O delineamento de parcelas subdivididas é similar ao de blocos aleatorizados (Figura 7.6), mas, nesse caso, um segundo fator de tratamento é aplicado ao bloco inteiro (= parcela).

Você deve comparar, com cuidado, o esquema de dois fatores (Figura 7.10) e o de parcelas subdivididas (Figura 7.12) para entender a sutil diferença entre eles. A distinção é que, no delineamento de dois fatores, cada réplica recebe as aplicações dos tratamentos independente e separadamente. No esquema de parcelas subdivididas, um dos tratamentos é aplicado a blocos ou parcelas inteiras e o outro tratamento, às réplicas dentro dos blocos.

A principal vantagem do delineamento de parcelas subdivididas é o uso eficiente de blocos para a aplicação de dois tratamentos. Como no delineamento de blocos aleatorizados, é um esquema simples que controla a heterogeneidade ambiental. Ele também pode ser menos trabalhoso do que aplicar tratamentos a réplicas individuais em um delineamento de dois fatores. O delineamento de parcelas subdivididas remove os efeitos aditivos dos blocos e permite testar os efeitos principais e as interações entre os dois fatores manipulados.[7]

Como no de blocos aleatorizados, o delineamento de parcelas subdivididas não permite testar a interação entre os blocos e o fator de subparcelas. No entanto, o delineamento de parcelas subdivididas permite testar o efeito principal do fator de parcela completa, o efeito principal do fator de subparcela e a interação entre os dois. Como nos aninhados, um erro comum é os investigadores analisarem um delineamento de parcelas subdivididas como se fosse uma ANOVA de dois fatores, o que aumenta o risco de erro Tipo I.

DELINEAMENTOS PARA TRÊS OU MAIS FATORES O delineamento de dois fatores pode ser estendido para três ou mais fatores. Por exemplo, se você estivesse estudando cascatas tróficas em uma cadeia alimentar de água doce (Brett & Goldman, 1997) poderia adicionar ou remover carnívoros de topo, predadores e herbívoros, e então medir os efeitos no nível dos produtores. Esse delineamento simples de três fatores gera $2^3 = 8$ combinações de tratamentos, incluindo uma que não tenha carnívoros de topo, predadores e nem herbívoros (Tabela 7.3). Como vimos, se fizermos um delineamento de blocos aleatorizados com um esquema de dois fatores replicando dentro de blocos, estes passam a ser um terceiro fator na análise. No entanto, delineamentos de três (ou mais) fatores raras vezes são usados em estudos ecológicos. Existem simplesmente muitas combinações de tratamentos para tornar esse delineamento logisticamente viável. Se

[7] Embora o exemplo que apresentamos tenha usado dois fatores manipulados experimentalmente, o delineamento de parcelas subdivididas também é eficiente quando um dos dois fatores representa uma fonte de variação natural. Por exemplo, em nossa pesquisa, estudamos a organização de cadeias tróficas aquáticas que se desenvolvem nas folhas cheias de água da chuva da planta-jarro *Sarracenia purpurea*. Um jarro novo se abre aproximadamente a cada 20 dias, fica cheio com água da chuva e rapidamente desenvolve uma cadeia alimentar associada de invertebrados e microrganismos.

Em um de nossos experimentos, manipulamos o regime de perturbação adicionando ou removendo água das folhas de cada planta (Gotelli et al., não publicado). Essas manipulações de água foram aplicadas a todos os jarros de uma planta. Depois, registramos a estrutura da cadeia alimentar no primeiro, no segundo e no terceiro jarros. Esses dados são analisados como um delineamento de parcelas subdivididas. O fator de parcela completa é o tratamento de água (5 níveis) e o fator de subparcela, a idade do jarro (3 níveis). A planta serviu como um bloco natural e foi eficiente e realístico aplicar os tratamentos de água a todas as folhas de uma planta.

TABELA 7.3 Combinações de tratamentos em um esquema com três fatores para um experimento de cadeia alimentar de adição e remoção

	Carnívoro ausente		Carnívoro presente	
	Herbívoro ausente	Herbívoro presente	Herbívoro ausente	Herbívoro presente
Produtor ausente	10	10	10	10
Produtor presente	10	10	10	10

Neste experimento, os três grupos tróficos representam os três fatores experimentais (carnívoros, herbívoros, produtores) cada um com dois níveis (ausente, presente). O valor em cada célula é o número de réplicas de cada combinação de tratamentos. Se a variável resposta for alguma propriedade da própria cadeia trófica, então a combinação de tratamentos em que todos os níveis tróficos estejam ausentes (na caixa) logicamente não será possível.

você achar que o seu está se tornando muito grande e complexo, deve considerar dividi-lo em um número menor de experimentos focados nas hipóteses principais que deseja testar.

INCORPORANDO A VARIABILIDADE TEMPORAL: DELINEAMENTOS DE MEDIDAS REPETIDAS Em todos os delineamentos que descrevemos até agora, a variável resposta é medida em cada réplica em um único ponto no tempo, no final do experimento. Um **delineamento de medidas repetidas** é usado sempre que observações múltiplas são coletadas na mesma réplica em tempos diversos. O delineamento de medidas repetidas pode ser visto como um de parcelas subdivididas em que uma única réplica atua como um bloco e o fator de subparcela é o tempo. Os delineamentos de medidas repetidas foram primeiramente usados em estudos de medicina e de psicologia nos quais repetidas observações eram realizadas por indivíduo. Assim, na terminologia de medidas repetidas, o **fator inter-sujeitos** corresponde ao de parcela completa, e o **fator intrassujeitos** corresponde aos diferentes tempos. No entanto, em um delineamento de medidas repetidas, as observações múltiplas em um único indivíduo não são independentes e as análises devem ser feitas com cautela.

Por exemplo, suponha que usamos um simples delineamento de um fator para o estudo das cracas, mostrado na Figura 7.5. Porém, além de fazer contagens em cada réplica apenas uma vez, medimos o número de cracas em cada réplica durante 4 semanas consecutivas. Agora, ao invés de 3 tratamentos × 10 réplicas = 30 observações, temos 3 tratamentos × 10 réplicas × 4 semanas = 120 observações (Tabela 7.4). Se usarmos dados de apenas uma das contagens, a análise poderia ser idêntica a um esquema de um fator.

Há três vantagens para o delineamento de medidas repetidas. A primeira é a eficiência: os dados são registrados em tempos diferentes, mas não é necessário ter réplicas individuais para cada combinação de tempo × tratamento. Segundo, o delineamento de medidas repetidas permite que cada réplica sirva como seu

TABELA 7.4 Planilha para uma análise de medidas repetidas simples

			Contagens de cracas			
Número ID	Tratamento	Réplica	Semana 1	Semana 2	Semana 3	Semana 4
1	Granito	1	12	15	17	17
2	Ardósia	1	10	6	19	32
3	Cimento	1	3	2	0	2
4	Granito	2	14	14	5	11
5	Ardósia	2	10	11	13	15
6	Cimento	2	8	9	4	4
7	Granito	3	11	13	22	29
8	Ardósia	3	11	17	28	15
9	Cimento	3	7	7	7	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	Cimento	10	8	0	0	3

Este experimento é delineado para testar os efeitos do tipo de substrato sobre o recrutamento de cracas na zona intertidal rochosa (Figura 7.5). O dados são organizados em uma planilha, na qual cada linha é uma réplica independente. As colunas indicam o número ID (1-30), o grupo de tratamento (cimento, ardósia, ou granito) e o número da réplica (1-10 dentro de cada tratamento). As quatro colunas seguintes fornecem o número de cracas registradas em um substrato particular em cada uma das quatro semanas consecutivas. As medidas, em diferentes momentos, não são independentes umas das outras, porque são obtidas em uma mesma réplica todas as semanas.

próprio bloco ou controle. Quando as réplicas representam indivíduos (plantas, animais, ou humanos), isto efetivamente controla a variação de tamanho, idade e história individual, que frequentemente exercem fortes influências na variável resposta. Por fim, o delineamento de medidas repetidas nos permite testar as interações do tempo com o tratamento. Por várias razões, esperamos que diferenças entre tratamentos possam mudar com o tempo. Em um experimento de pressão (Capítulo 6), pode haver efeitos cumulativos do tratamento que não serão expressos até algum tempo depois do início do experimento. Em contraste, em um experimento de pulso, esperamos ver as diferenças entre tratamentos diminuírem à medida que o tempo passa desde o único pulso de aplicação do tratamento. Tais efeitos complexos são melhor observados na interação entre tempo e tratamento, e podem não ser detectados se a variável resposta for medida em um único ponto no tempo.

O delineamento de blocos aleatorizados e o de medidas repetidas assumem um pressuposto especial de **circularidade** para o fator intrassujeitos. Circularidade (no contexto de ANOVA) significa que a variância das diferenças entre quaisquer dois níveis de tratamento na subparcela é a mesma. Para o delineamento de blocos aleatorizados, isso significa que a variância da diferença entre qualquer par de tratamentos no bloco é a mesma. Se as parcelas dos tratamentos

forem grandes o bastante além de adequadamente espaçadas, este frequentemente será um pressuposto razoável. Para o delineamento de medidas repetidas, o pressuposto de circularidade significa que a variância da diferença das observações entre qualquer par de tempos é a mesma. O pressuposto de circularidade é improvável de ser alcançado para medidas repetidas. Na maioria dos casos, a variância da diferença entre duas observações consecutivas é provavelmente muito menor que a variabilidade da diferença entre duas observações que são amplamente espaçadas no tempo. Isso porque é provável que séries temporais medidas sobre um mesmo sujeito apresentem uma "memória" temporal, de forma que os valores atuais sejam uma função dos valores observados no passado recente. Esta premissa de observações correlacionadas é a base para análises de séries temporais (Capítulo 6).

A desvantagem principal com análises de medidas repetidas é a falha em cumprir o pressuposto de circularidade. Se as medidas repetidas são serialmente correlacionadas, as taxas de erro Tipo I para testes F serão infladas e a hipótese nula pode ser rejeitada incorretamente quando ela for verdadeira. A melhor maneira de cumprir o pressuposto de circularidade é usar tempos amostrais espaçados com regularidade em conjunto ao conhecimento de história natural de seus organismos para selecionar um intervalo amostral apropriado.

Para análise de medidas repetidas que não dependam do pressuposto de circularidade, há algumas alternativas: uma abordagem é estabelecer réplicas suficientes para que um conjunto diferente seja amostrado em cada período de tempo. Com esse delineamento, o tempo pode ser tratado como um fator simples em uma análise de variância de dois fatores. Se os métodos amostrais são destrutivos (p. ex., coletar o conteúdo estomacal de peixes, matar e preservar uma amostra de invertebrados, ou arrancar plantas), esta é a única maneira de incorporar o tempo ao delineamento.

Uma segunda estratégia é usar um esquema de medidas repetidas, mas ser mais criativo no delineamento da variável resposta. Unir as medidas repetidas correlacionadas em uma única variável resposta para cada indivíduo e então usar uma simples análise de variância de um fator. Por exemplo, se você deseja testar se tendências temporais diferem entre tratamentos (o fator inter-sujeitos), você poderia ajustar uma linha de regressão (com um modelo linear ou um de série temporal) aos dados de medidas repetidas e usar a inclinação da linha como variável resposta. Um valor de inclinação separado poderia ser calculado para cada um dos indivíduos no estudo. As inclinações poderiam, então, ser comparadas usando uma análise simples de um fator, tratando cada indivíduo como uma observação independente (e são). Os efeitos significativos do tratamento poderiam indicar diferentes trajetórias temporais para indivíduos nos diversos tratamentos. Esse teste é muito parecido ao da interação entre tempo e tratamento em uma análise de medidas repetidas padrão.

Embora tais variáveis compostas sejam criadas a partir de observações correlacionadas coletadas a partir de um indivíduo, elas são independentes entre indivíduos. Além disso, o Teorema do Limite Central (Capítulo 2) nos diz que as médias desses valores vão seguir uma distribuição aproximadamente nor-

mal mesmo se as próprias variáveis subjacentes não sigam. Como a maioria dos dados de medidas repetidas não cumpre o pressuposto de circularidade, nosso conselho é ser cuidadoso com tais análises. Preferimos unir os dados temporais em uma única variável que seja verdadeiramente independente entre observações e, então, usar um delineamento mais simples de um fator para a análise.

IMPACTOS AMBIENTAIS AO LONGO DO TEMPO: DELINEAMENTOS ADCI Um tipo especial de medidas repetidas é um em que são realizadas medidas antes e depois da aplicação de um tratamento. Por exemplo, suponha que você esteja interessado em medir o efeito da atrazina (um composto que mimetiza hormônios) sobre o tamanho corporal de sapos (Allran & Karasov, 2001). Em um esquema simples de um fator, você poderia atribuir os sapos de maneira aleatória a grupos controle e tratamento, aplicar a atrazina e medir o tamanho corporal ao final do experimento. Um delineamento mais sensível seria estabelecer os grupos controle e tratamento e fazer medidas de tamanho corporal durante um ou mais períodos de tempo antes da aplicação do tratamento. Depois, você mede novamente o tamanho corporal várias vezes nos grupos controle e tratamento.

Esses delineamentos também são usados em estudos observacionais que avaliam impactos ambientais. Em avaliação de impactos, as medidas são feitas antes e depois de ele acontecer. Uma avaliação típica poderia ser das potenciais respostas de uma comunidade de invertebrados marinhos à operação de uma usina de energia nuclear, que despeja um considerável efluente de água quente (Schroeter et al., 1993). Antes de a usina iniciar a operação, você faz uma ou mais amostras na área que será afetada e estima a abundância de espécies de interesse (p. ex., caramujos, estrelas-do-mar, ouriços-do-mar). A replicação neste estudo pode ser espacial, temporal ou ambas. A primeira exigiria amostras em muitas parcelas diferentes dentro e fora da projeção da mancha de água quente despejada.[8] A replicação temporal exigiria amostrar um único local na área de despejo em vários momentos antes de a usina ser ativada. Idealmente, múltiplos locais poderiam ser amostrados várias vezes no período pré-despejo.

Uma vez que o despejo tenha começado, o protocolo de amostragem é repetido. Nesse delineamento de avaliação, é indispensável que haja no mínimo um local de controle ou de referência que seja amostrado ao mesmo tempo, tanto antes quanto depois do despejo. Então, se você observar um declínio na abundância no local impactado, mas não no local de controle, você pode testar quando o declínio será significativo. Além disso, a abundância de invertebrados pode estar diminuindo por razões que nada têm a ver com o despejo de água quente. Neste

[8] Um pressuposto-chave neste esquema é que o investigador conhece por antecedência o alcance espacial do impacto. Sem esta informação, algumas das parcelas "controle" podem terminar dentro da região do "tratamento" e o efeito da mancha de água quente poderia ser subestimado.

caso, você encontraria menor abundância tanto nos locais de controle quanto nos impactados.

Este tipo de delineamento de medidas repetidas é tratado como um **delineamento ADCI** (Antes-Depois, Controle-Impacto). Não há apenas replicação de parcelas-controle e tratamento, existe também replicação temporal com medidas anteriores e posteriores à aplicação do tratamento (Figura 7.13).

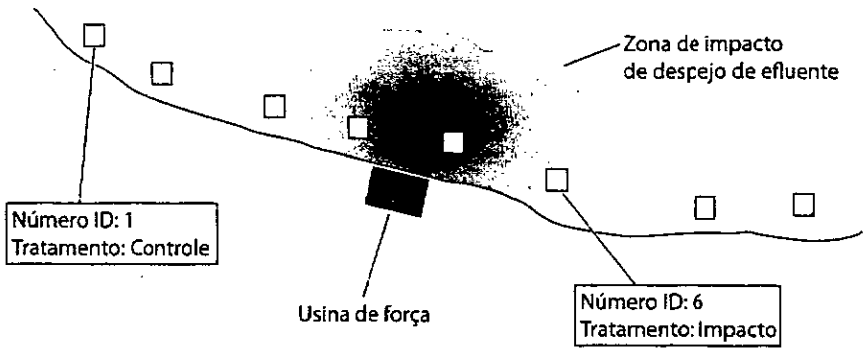
Em sua forma ideal, o delineamento ADCI é um esquema poderoso para avaliar perturbações ambientais e monitorar trajetórias antes e depois do impacto. A replicação no espaço garante que os resultados serão aplicáveis a outros locais que possam vir a ser perturbados da mesma forma. A replicação ao longo do tempo garante que a trajetória temporal de resposta e recuperação possa ser monitorada. O delineamento é apropriado tanto para distúrbios ou experimentos de pulso quanto para experimentos de pressão.

Infelizmente, este delineamento ADCI idealizado raras vezes é alcançado em estudos de impacto ambiental. Muitas vezes existe apenas um único local, que será impactado, e em geral não é escolhido aleatoriamente (Stewart-Oaten & Bence 2001, Murtaugh, 2002b). As réplicas espaciais dentro da área impactada não são independentes, porque existe apenas um único impacto estudado em um único local (Underwood, 1994). Se o impacto representa um acidente ambiental, como um derramamento de óleo, pode não haver nenhum dado pré-impacto disponível, nem para locais de referência quanto para os impactados.

O controle potencial da aleatorização e atribuição de tratamentos são muito melhores em manipulações experimentais em larga escala, mas mesmo nestes casos haverá pouca, se houver alguma, replicação espacial. Tais estudos dependem de replicação temporal mais intensiva, tanto antes quanto depois da manipulação. Por exemplo, desde 1983, Brezonik et al. (1986) conduzem um experimento de acidificação de longo prazo no Lago Little Rock, um pequeno lago oligotrófico de infiltração no norte de Wisconsin. O lago foi dividido em duas represas, uma de tratamento e uma de referência, usando uma lona de vinil impermeável. Dados de referência (pré-manipulação) foram coletados em ambas de agosto de 1983 até abril de 1985. A represa de tratamento foi então acidificada, de maneira gradativa, com ácido sulfúrico em três níveis de pH (5,6; 5,1; 4,7). Esses níveis de pH foram mantidos em um experimento de pressão com intervalos de dois anos.

Como havia apenas uma bacia de tratamento e uma bacia de controle, os métodos convencionais de ANOVA não podem ser usados para analisar esses dados.[9] Existem duas estratégias gerais para análise. A **análise de intervenção randomizada** (RIA) é um procedimento de Monte Carlo (*ver* Capítulo 5), no qual uma

[9] Com certeza, se alguém assumir que as amostras são réplicas independentes, a ANOVA convencional pode ser usada. Mas não há prudência em forçar os dados a uma estrutura de modelo à qual eles não se ajustam. Um dos assuntos centrais deste capítulo é escolher delineamentos simples para os seus experimentos e amostragens, cujos pressupostos se adaptem melhor às restrições dos seus dados.



Número ID	Tratamento	Amostragem pré-impacto				Amostragem pós-impacto			
		Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8
1	Controle	106	108	108	120	122	123	130	190
2	Controle	104	88	84	104	106	119	135	120
3	Impacto	99	97	102	192	150	140	145	150
4	Impacto	120	122	98	120	137	135	155	165
5	Impacto	88	90	92	94	0	7	75	77
6	Impacto	100	120	129	82	2	3	66	130
7	Controle	66	70	70	99	45	55	55	109
8	Controle	130	209	220	250	100	90	88	140

Figura 7.13 Exemplo do arranjo espacial de réplicas para um delineamento ADCI. Cada quadrado representa uma parcela amostral em uma região costeira que será potencialmente afetada pelo despejo de água quente de uma usina nuclear. Parcelas permanentes são estabelecidas dentro da zona do efluente de água quente (área sombreada) e em zonas controle adjacentes (áreas não sombreadas). Todas as parcelas são amostradas semanalmente, durante 4 semanas, antes de a usina iniciar o despejo água quente e por 4 semanas após. Cada linha da planilha representa uma réplica diferente. Duas colunas indicam o número ID da réplica e o tratamento (Controle ou Impacto). As colunas restantes dão os dados de abundância de invertebrados coletados em cada uma das 8 datas de amostragem (4 datas de amostragem pré-despejo, 4 datas de amostragem pós-despejo).

estatística de teste calculada a partir da série temporal é comparada a uma distribuição de valores criada pela aleatorização ou permutação dos dados da série temporal entre os intervalos de tratamento (Carpenter et al., 1989). A RIA relaxa o pressuposto de normalidade, mas ela continua suscetível às correlações temporais nos dados (Stewart-Oaten et al., 1992).

Uma segunda estratégia é usar a análise de séries temporais para ajustar modelos simples aos dados. O modelo **autorregressivo integrado de médias móveis** (ARIMA) descreve a estrutura de correlação em dados temporais com poucos parâmetros (*ver* Capítulo 6). Os parâmetros adicionais do modelo estimam as mudanças gradativas que ocorrem com as intervenções experimentais e esses parâmetros podem ser testados contra a hipótese nula de que eles não diferem de 0. Os modelos ARIMA podem ser ajustados individualmente aos dados da série temporal de controle e de manipulação, ou a uma série derivada de dados criada

pela razão entre os dados de tratamento/controle de cada passo de tempo (Rasmussen et al., 1993). Os métodos bayesianos também podem ser usados para analisar dados de delineamentos ADCI (Carpenter et al., 1996; Rao & Tirtotjondro, 1996; Reckhow, 1996; Varis & Kuikka, 1997; Fox, 2001).

A RIA, o ARIMA e os métodos bayesianos são ferramentas poderosas para detectar efeitos de tratamentos em dados de séries temporais. No entanto, sem replicação, ainda será problemático generalizar os resultados da análise. O que poderia acontecer em outros lagos? Outros anos? Informações adicionais, incluindo os resultados de experimentos em pequena escala (Frost et al., 1988), ou comparações instantâneas com um grande número de locais de controle sem manipulação (Schindler et al., 1985; Underwood, 1994) podem ajudar a expandir o domínio de inferência a partir de um estudo ADCI.

Alternativas à ANOVA: experimento de regressão

A literatura sobre delineamento experimental é dominada por esquemas de ANOVA e a ciência ecológica moderna tem sido tratada sarcasticamente como pouco mais que o cuidado e curadoria de tabelas de ANOVA. Embora os delineamentos de ANOVA sejam convenientes e poderosos para vários propósitos, nem sempre são a melhor escolha. A ANOVA se tornou tão popular que pode agir como uma camisa-de-força intelectual (Werner, 1998) e levar os cientistas a negligenciar outros delineamentos experimentais úteis.

Sugerimos que um delineamento de regressão seja aplicado, por parecer mais propriado, em vez dos delineamentos de ANOVA. Em muitos delineamentos de ANOVA uma variável preditora contínua é testada com apenas alguns poucos valores, de forma que ela possa ser tratada como uma variável preditora categórica e ser condensada em um delineamento como esse. Exemplos incluem níveis de tratamento que representam diferentes concentrações de nutrientes, temperaturas ou níveis de recursos.

Em contraste, um delineamento experimental de regressão (Figura 7.14) usa vários níveis diferentes da variável contínua independente; depois, usa regressão para ajustar uma linha, uma curva ou uma superfície aos dados. Uma questão complicada em tal delineamento (o mesmo problema está presente em uma ANOVA) é escolher os níveis apropriados para a variável preditora. Uma seleção uniforme de valores preditores dentro da amplitude desejada deve garantir maior poder estatístico e um ajuste razoável à linha de regressão. No entanto, se a resposta esperada for multiplicativa, em vez de linear (p. ex., uma redução de 10% no crescimento a cada duplicação da concentração), pode ser melhor estabelecer os valores do preditor em uma escala logarítmica espaçada de maneira uniforme. Nesse delineamento você terá mais dados coletados em baixas concentrações, no qual se espera que as mudanças na variável resposta sejam mais abruptas.

Uma das principais vantagens dos delineamentos de regressão é a eficiência. Suponha que você esteja estudando respostas de plantas terrestres e comunida-

(Tratamentos simultâneos de N e P em um delineamento de ANOVA de dois fatores)

		Tratamento de nitrogênio	
		0,00 mg	0,50 mg
Tratamento de fósforo	0,00 mg	12	12
	0,05 mg	12	12

(Delineamento experimental de regressão com dois fatores)

		Tratamento de nitrogênio						
		0,00 mg	0,05 mg	0,10 mg	0,20 mg	0,40 mg	0,80 mg	1,00 mg
Tratamento de fósforo	0,00 mg	1	1	1	1	1	1	1
	0,01 mg	1	1	1	1	1	1	1
	0,05 mg	1	1	1	1	1	1	1
	0,10 mg	1	1	1	1	1	1	1
	0,20 mg	1	1	1	1	1	1	1
	0,40 mg	1	1	1	1	1	1	1
	0,50 mg	1	1	1	1	1	1	1

Figura 7.14 Combinação de tratamentos em um delineamento de ANOVA de dois fatores (painel superior) e um delineamento experimental de regressão (painel inferior). Esses experimentos testam os efeitos aditivos e de interação do nitrogênio (N) e do fósforo (P) sobre o crescimento de plantas ou alguma outra variável resposta. Cada célula, na tabela, indica o número de réplicas usadas. Se o número máximo de réplicas for 50 e um mínimo de 10 por tratamentos seja necessário, apenas 2 níveis de tratamento são possíveis para cada N e P (cada um com 12 réplicas) na ANOVA de dois fatores. Em contraste, a regressão experimental permite 7 níveis de tratamentos para cada N e P. Cada uma das $7 \times 7 = 49$ parcelas no delineamento recebe uma combinação única de concentrações de N e P.

des de insetos ao nitrogênio (N), e que seu tamanho amostral total seja limitado a 50 parcelas, por causa do espaço ou esforço disponível. Se você tentar seguir a Regra dos 10, um delineamento de ANOVA poderia forçá-lo a selecionar apenas 5 níveis diferentes de fertilização e replicar cada um 10 vezes. Embora esse delineamento seja adequado para alguns propósitos, ele pode não ajudar a identificar os níveis críticos limiares nos quais a estrutura da comunidade muda dramaticamente em resposta ao N. Em contraste, um delineamento de regressão permitiria você estabelecer 50 níveis diferentes de N, um por parcela. Com esse delineamento você pode caracterizar, com acurácia, as mudanças que ocorram na estrutura da comunidade com o aumento dos níveis de N. As exibições gráficas podem ajudar a revelar os pontos limiares e efeitos não lineares (*ver* Capítulo 9). Certamente, até mesmo a replicação mínima de cada nível de tratamento é muito desejável, mas, se o tamanho amostral total for limitado, pode não ser possível.

Para uma ANOVA de dois fatores, um experimento de regressão é ainda mais eficiente e poderoso. Se você deseja manipular nitrogênio e fósforo (P) como fatores independentes e ainda manter 10 réplicas por tratamento, não poderia ter mais do que dois níveis de N e dois de P. Como um dos níveis precisa ser uma parcela-controle (i. e., sem fertilização), o experimento não lhe fornecerá muita informação sobre o papel dos diferentes níveis de N e P no sistema. Se o resulta-

do for estatisticamente significativo, você poderá dizer apenas que a comunidade responde àqueles níveis de N e P em particular, que é algo que você já poderia saber a partir da literatura antes de ter começado. Se, ao contrário, o resultado não for estatisticamente significativo, a crítica óbvia é que as concentrações de N e P foram muito baixas para gerar uma resposta.

Em contraste, um delineamento experimental de regressão poderia ser um delineamento completamente cruzado com 7 níveis de nitrogênio e 7 níveis de fósforo, com um nível de cada correspondendo ao controle (sem N e sem P). Cada uma das $7 \times 7 = 49$ réplicas recebe uma concentração única de N e P, e uma das 49 parcelas não recebe N nem P (Figura 7.14). Eis um delineamento de superfície de resposta (Inouye, 2001), no qual a variável resposta será modelada por regressão múltipla. Com 7 níveis de cada nutriente, esse delineamento proporciona um teste muito mais robusto para efeitos aditivos e de interação dos nutrientes, e ainda poderia revelar respostas não lineares. Se os efeitos de N e P são fracos ou sutis, o modelo de regressão terá mais chance de revelar efeitos significativos do que a ANOVA de dois fatores.[10]

A eficiência não é a única vantagem de um delineamento experimental de regressão. Ao representar naturalmente a variável preditora em uma escala contínua, será muito mais fácil detectar respostas não lineares, limiares ou assintóticas. Isso não pode ser inferido com confiança a partir de um modelo de ANOVA, que em geral não tem níveis de tratamentos suficientes para ser informativo. Se a relação for outra que não uma linha reta, existe uma gama de métodos estatísticos para ajustar respostas não lineares (*ver* Capítulo 9).

Uma última vantagem em usar um delineamento experimental de regressão é o benefício potencial de integrar seus resultados com predições teóricas e modelos ecológicos. A ANOVA fornece estimativas de médias e variâncias para grupos ou níveis de variáveis categóricas em particular. Essas estimativas raras vezes são de interesse ou uso em modelos ecológicos. Em contraste, uma análise de regressão fornece estimativas dos parâmetros de inclinação e intercepto que medem a mudança na resposta Y relativa à mudança no preditor X (dY/dX). Essas derivadas são o que se precisa para testar vários modelos ecológicos descritos como simples equações diferenciais.

[10] Há, ainda, uma penalização oculta que geralmente não é reconhecida em usar um delineamento de ANOVA de dois fatores para este experimento. Se os níveis do tratamento representam um pequeno subconjunto de vários outros níveis possíveis, então o delineamento é tratado como um modelo de **ANOVA de efeitos aleatórios**. A menos que exista algo especial acerca dos níveis dos tratamentos em particular que foram usados, um modelo de efeitos aleatórios sempre é a escolha mais apropriada quando uma variável contínua tenha sido condensada em uma categórica para ANOVA. No modelo de efeitos aleatórios, o denominador no teste da razão-F para os efeitos do tratamento é o quadrado médio da interação, e não o quadrado médio do erro, que é usado no modelo padrão de **ANOVA de efeitos fixos**. Se não houver muitos níveis de tratamento, não haverá muitos graus de liberdade associados com o termo de interação, independente da quantidade de replicação dentro dos tratamentos. *Ver* Capítulo 10 para mais detalhes sobre modelos de ANOVA com efeitos fixos e aleatórios e a para a construção das razões-F.

Uma abordagem de regressão experimental pode não ser viável se for muito caro ou demorado estabelecer níveis únicos da variável preditora. Nesse caso, um delineamento de ANOVA pode ser preferível, porque apenas alguns poucos níveis da variável preditora podem ser estabelecidos. Uma desvantagem aparente do experimento de regressão é que ele aparenta não ter replicação! Na Figura 7.14, cada nível único de tratamento é aplicado em apenas uma réplica e isso parece ir contra o princípio de replicação (*ver* Capítulo 6). Se cada tratamento único não for replicado, a solução de mínimos quadrados para a linha de regressão ainda fornece uma estimativa dos parâmetros da regressão e de suas variâncias (*ver* Capítulo 9). A linha de regressão fornece uma estimativa não enviesada do valor esperado da resposta Y para um dado valor do preditor X e a estimativa da variância pode ser usada para construir um intervalo de confiança acerca dessas estimativas. Isso, na verdade, é mais informativo que os resultados de um modelo de ANOVA, que lhe permite estimar médias e intervalos de confiança para apenas o punhado de níveis de tratamentos que foram usados.

Uma questão potencialmente mais séria é que o delineamento de regressão pode não incluir qualquer replicação do controle. Em nosso exemplo de dois fatores, há apenas uma parcela sem adição de nitrogênio e fósforo. Se esse será um problema sério ou não dependerá dos detalhes do delineamento experimental. Uma vez que todas as réplicas sejam tratadas igualmente e difiram apenas na aplicação do tratamento, o experimento permanece válido e os resultados estimarão, com acurácia, os efeitos relativos dos diferentes níveis da variável preditora sobre a variável resposta. Se for desejável estimar o efeito absoluto do tratamento, então um adicional de réplicas de parcelas-controle pode ser necessário para levar em conta qualquer efeito do manuseio ou outras respostas às condições experimentais em geral. Essas questões não são diferentes daquelas encontradas em delineamentos de ANOVA.

Historicamente, a regressão tem sido utilizada de forma predominante na análise de dados não experimentais, muito embora seus pressupostos sejam improváveis de serem cumpridos na maioria dos estudos amostrais. Um estudo experimental com base em um delineamento de regressão não apenas cumpre os pressupostos da análise, mas com frequência também é mais robusto e apropriado que um delineamento de ANOVA. Aconselhamos a "pensar fora da caixinha da ANOVA" e considerar um delineamento de regressão quando estiver manipulando variáveis preditoras contínuas.

Delineamentos tabulares

A última classe de delineamentos experimentais é usada quando a variável preditora e a resposta são categóricas. As medidas nesse delineamento são contagens. A variável mais simples é uma resposta dicotômica (ou binomial, *ver* Capítulo 2) em uma série de tentativas independentes. Por exemplo, em um teste sobre o comportamento de baratas, você pode colocar uma barata em uma arena com um lado preto e outro branco e registrar em que lado o animal passa a maior

parte do tempo. Para garantir independência, cada barata replicada seria testada individualmente.

Na maioria das vezes, uma resposta dicotômica será registrada para duas ou mais categorias da variável preditora. No estudo das baratas, metade delas poderia estar infectada experimentalmente com um parasita conhecido por alterar o comportamento do hospedeiro (Moore, 1984). Agora questionamos se a resposta das baratas difere entre indivíduos parasitados e não parasitados (Moore, 2001; Poulin, 2000). Essa abordagem poderia ser estendida para um delineamento de três fatores, adicionando um tratamento adicional e perguntando se a diferença entre indivíduos parasitados e não parasitados muda na presença ou ausência de um vertebrado predador.

Poderíamos predizer que é mais provável que indivíduos não infectados usem o substrato preto, que irá torná-lo menos evidente a um predador visual. Na presença do predador, os indivíduos não infectados se deslocariam ainda mais para as superfícies escuras, enquanto os infectados, para as superfícies brancas. Além disso, o parasita poderia alterar o comportamento do hospedeiro, mas essas alterações poderiam ser independentes da presença ou ausência do predador. Ainda outra possibilidade seria a de que o comportamento do hospedeiro poderia ser muito sensível à presença do predador, mas não necessariamente afetado pela infecção com o parasita. Uma **análise de tabela de contingência** (Capítulo 11) é usada para testar todas essas hipóteses com o mesmo conjunto de dados.

Em alguns delineamentos tabulares, o investigador determina o número total de indivíduos em cada categoria da variável preditora e esses indivíduos serão classificados de acordo com suas repostas. O total para cada categoria é tratado como **total marginal,** porque representa a soma da coluna ou da linha na margem da tabela de dados. Em um estudo observacional, o investigador pode determinar um ou ambos dos totais marginais ou apenas o total geral das observações independentes. Em um delineamento tabular, o total geral é igual à soma do total marginal da coluna ou da linha.

Digamos que você esteja tentando determinar as associações de quatro espécies de lagartos *Anolis* com três tipos de micro-hábitats (solo, troncos de árvores, galhos de árvores; *ver* Butler & Losos, 2002). A Tabela 7.5 mostra o esquema de dois fatores dos dados desse estudo. Cada linha na tabela representa uma espécie diversa de lagarto e cada coluna representa uma categoria de hábitat diferente. Os valores em cada célula representam as contagens de uma espécie de lagarto em especial registradas em um hábitat em particular. O total marginal das linhas representa o número total de observações para cada espécie de lagarto, somadas ao longo dos três tipos de hábitats. O total marginal das colunas representa o número total de observações em cada tipo de hábitat, somadas ao longo das quatro espécies de lagarto. O total geral na tabela (N = 81) representa a contagem total de todas as espécies de lagarto observadas em todos os hábitats.

TABELA 7.5 Contagem tabular da ocorrência de quatro espécies de lagartos contadas em três micro-hábitats diferentes

		Hábitat			
		Solo	Tronco de árvore	Galho de árvore	Total das espécies
Espécies de lagarto	Espécie A	9	0	15	*24*
	Espécie B	9	0	12	*21*
	Espécie C	9	5	0	*14*
	Espécie D	9	10	3	*22*
	Total dos Hábitats	36	15	30	81

Os valores em itálico são os totais marginais para a tabela de dois fatores. O tamanho amostral total é de 81 observações. Nestes dados, tanto a variável resposta (identidade das espécies) quanto a variável preditora (categoria de micro-hábitat) são categóricas.

Estes dados poderiam ter sido coletados de diversas maneiras, dependendo se a amostragem foi com base no total marginal para o micro-hábitat, no total marginal para os lagartos ou no total geral para a amostra inteira.

Em um esquema amostral construído acerca de micro-hábitats, o investigador poderia gastar 10 horas amostrando cada um e registrando o número de espécies diferentes de lagartos encontradas em cada hábitat. Assim, na contagem de troncos de árvores, o investigador encontrou um total de 15 lagartos: 5 da Espécie C e 10 da Espécie D, as Espécies A e B não foram encontradas. Na contagem no solo, o investigador encontrou 36 lagartos com todas as espécies igualmente representadas.

Como alternativa, a amostragem poderia ter sido fundamentada nos próprios lagartos. Em tal delineamento, o investigador deveria determinar um esforço amostral igual para cada espécie, procurando todos os hábitats de maneira aleatória, em busca de indivíduos de uma espécie de lagarto em especial e registrando em qual hábitat ela ocorreu. Assim, em uma busca pela Espécie B, 21 indivíduos foram encontrados, 9 no solo e 12 em galhos de árvores. Outra variante amostral é uma na qual os totais de linha e de coluna são simultaneamente fixados. Embora esse delineamento não seja muito comum em estudos ecológicos, ele pode ser usado para um teste estatístico exato da distribuição dos valores amostrados (Teste Exato de Fisher; *ver* Capítulo 11). Por fim, a amostragem poderia basear-se apenas no total das observações. Assim, o investigador poderia ter feito uma amostra aleatória de 81 lagartos e para cada encontrado ter registrado a identidade da espécie e do micro-hábitat.

Idealmente, os totais marginais nos quais a amostragem é fundamentada deveriam ser os mesmos para cada categoria, assim como tentamos alcançar tamanhos amostrais iguais ao estabelecer um delineamento de ANOVA. No entanto, tamanhos amostrais idênticos não são necessários na análise de dados tabulares. Entretanto, os testes requerem (como sempre) que as observações

sejam amostradas de maneira aleatória e que as réplicas sejam verdadeiramente independentes umas das outras. Isso pode ser muito difícil de obter-se em alguns casos. Por exemplo, se os lagartos tendem a se agregar ou a se mover em grupo, não podemos simplesmente contar indivíduos conforme são encontrados, pois é provável que um grupo inteiro seja encontrado em um mesmo micro-hábitat. Nesse exemplo, também assumimos que todos os lagartos sejam igualmente visíveis ao observador em todos os hábitats. Se algumas espécies ficarem mais visualizáveis em certos hábitats que outras, então as frequências relativas irão refletir vieses amostrais em vez de associações das espécies com os micro-hábitats.

DELINEAMENTOS AMOSTRAIS PARA DADOS CATEGÓRICOS Em contraste à extensa literatura para delineamentos de regressão e ANOVA, relativamente pouco tem sido escrito, em um contexto ecológico, sobre delineamentos amostrais para dados categóricos. Se as observações são caras e demoradas, todo esforço deve ser feito para garantir que cada observação seja independente, de forma que um simples esquema de dois ou múltiplos fatores possa ser usado. Infelizmente, várias análises de dados categóricos publicadas são fundamentadas em observações não independentes, algumas coletadas em diferentes momentos ou lugares. Muitos estudos comportamentais analisam múltiplas observações de um mesmo indivíduo. Tais dados claramente não podem ser tratados como independentes (Kramer & Schmidhammer, 1992). Se os dados tabulares não são amostras aleatórias independentes do mesmo espaço amostral, você deveria incorporar as categorias temporais ou espaciais como fatores em sua análise.

Alternativas para os delineamentos tabulares: delineamentos proporcionais

Se as observações individuais forem acessíveis e puderem ser obtidas em grande quantidade, existe uma alternativa aos delineamentos tabulares: uma das variáveis categóricas pode ser transformada em uma medida de proporção (número de resultados desejados/número de observações), que é uma variável contínua. A variável contínua pode, então, ser analisada usando qualquer um dos métodos descritos acima por regressão ou ANOVA.

Há duas vantagens de usar delineamentos proporcionais em vez de tabulares. A primeira é que o conjunto padrão de delineamentos de ANOVA e regressão pode ser usado, incluindo o uso de blocos. A segunda vantagem é que a análise de proporções pode ser usada para acomodar dados de frequência que não são estritamente independentes. Por exemplo, suponha que, para ganhar tempo, o experimento das baratas foi estabelecido com 10 indivíduos colocados na arena de uma só vez. Não seria legítimo tratar os 10 indivíduos como réplicas independentes, pela mesma razão que as subamostras de uma única gaiola não são réplicas independentes em uma ANOVA. No entanto, os dados para esta execução podem ser tratados como uma única réplica, para a qual podemos calcular a proporção de indivíduos presentes no lado escuro da arena. Com múltiplas execuções do experimento, agora podemos testar hipóteses acerca das diferenças

na proporção entre grupos (p. ex., parasitados *versus* não parasitados). O delineamento ainda é problemático, pois é possível que a seleção de substratos por baratas solitárias seja diferente das de substratos por grupos de baratas. Todavia, o delineamento ao menos evita tratar indivíduos dentro da arena como réplicas independentes, o que não são.

Embora as proporções, como as probabilidades, sejam variáveis contínuas, elas estão delimitadas entre 0,0 e 1,0. A transformação do arcoseno da raiz quadrada dos dados de proporção pode ser necessária para cumprir o pressuposto de normalidade (Capítulo 8). Uma segunda consideração na análise de dados proporcionais é a importância em usar pelo menos 10 tentativas por réplica e certificar-se de que os tamanhos amostrais sejam os mais equilibrados possível. Com 10 indivíduos por réplica, as possíveis medidas da variável resposta estão no conjunto {0,0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1,0}. Mas suponha que o mesmo tratamento tenha sido aplicado a uma réplica na qual apenas três indivíduos foram usados. Neste caso, os únicos valores possíveis estão no conjunto {0,0; 0,33; 0,66; 1,0}. Este pequeno tamanho amostral inflará muito a variância medida e tal problema não é aliviado por qualquer transformação de dados.

Um problema final com a análise de proporções surge se existirem três ou mais variáveis categóricas. Com uma resposta dicotômica, a proporção caracteriza completamente os dados. No entanto, se houver mais de duas categorias, a proporção deverá ser definida com cuidado em termos de apenas uma das categorias. Por exemplo, se a arena incluir superfícies pretas e brancas verticais e horizontais, haverá quatro categorias a partir das quais as proporções podem ser medidas. Assim, a análise poderia ser com base na proporção de indivíduos usando a superfície preta horizontal. Alternativamente, a proporção poderia ser definida em termos de duas ou mais categorias somadas, como a proporção de indivíduos usando qualquer superfície vertical (branca ou preta).

RESUMO

As variáveis dependentes e independentes são categóricas ou contínuas, e a maioria dos delineamentos se ajusta em uma entre quatro categorias possíveis com base nesta classificação. Os delineamentos de análise de variância (ANOVA) são usados para experimentos cuja variável independente é categórica e a dependente é contínua. Os delineamentos úteis de ANOVA incluem as ANOVAs de um e dois fatores, blocos aleatorizados e delineamentos de parcelas subdivididas. Não apoiamos o uso de ANOVAs aninhadas, na qual subamostras não independentes são realizadas dentro de uma réplica. Os delineamentos de medidas repetidas podem ser usados quando observações repetidas são coletadas em uma única réplica ao longo do tempo. No entanto, esses dados em geral são autocorrelacionados, de forma que os pressupostos da análise podem não ser cumpridos. Nesses casos, os dados temporais deveriam ser condensados em uma única medida independente ou em uma análise de séries temporais deveria ser usada.

Se a variável independente é contínua, um delineamento de regressão é usado. Os delineamentos de regressão são apropriados tanto para estudos experimentais quanto observacionais, embora eles sejam usados predominantemente no último. Defendemos o uso crescente de experimentos de regressão em vez de ANOVA com apenas alguns poucos níveis da variável independente representados. A amostragem adequada da gama de valores do preditor é importante no delineamento de um experimento de regressão adequado. Os delineamentos de regressão múltipla incluem duas ou mais variáveis preditoras, embora a análise se torne problemática caso haja correlações fortes (colinearidade) entre as variáveis preditoras.

Se tanto a variável independente quanto a dependente são categóricas, um delineamento tabular é empregado. Os delineamentos tabulares requerem réplicas de contagem verdadeiramente independentes. Se as contagens não são independentes elas devem ser transformadas de forma que a variável resposta seja uma simples proporção. O delineamento experimental será então similar a uma regressão ou ANOVA.

Defendemos delineamentos amostrais e experimentais simples e enfatizamos a importância de coletar dados de réplicas que são independentes umas das outras. Boa replicação e tamanhos amostrais balanceados aumentarão o poder e a confiabilidade das análises. Mesmo as mais sofisticadas não podem salvar os resultados de um estudo maldelineado.

8

Gestão e Curadoria de Dados

O conjunto de dados é a matéria bruta dos estudos científicos. No entanto, raras vezes dedicamos a eles a mesma quantidade de tempo e energia na organização, gestão e curadoria de que disponibilizamos na coleta, análise e publicação dos dados. Infelizmente, é muito mais fácil usar os próprios dados originais (brutos) para testar novas hipóteses que extraí-los da literatura. Um requisito básico de qualquer estudo científico é que ele possa ser repetido por outros pesquisadores. As análises não podem ser reconstruídas, a menos que os dados originais sejam devidamente organizados e catalogados, armazenados em segurança e tornados disponíveis. Além disso, é uma exigência legal que os dados coletados como parte de um projeto financiado por recursos públicos (p. ex., subvenções e contratos federais ou estaduais) sejam disponibilizados a outros cientistas e para o público. O compartilhamento de dados é legalmente exigido, pois, tecnicamente, a agência financiadora é dona dos dados, e não o pesquisador. A maioria das agências financiadoras concede um bom tempo (em geral, um ou mais anos) para os pesquisadores comunicarem seus resultados, antes que os dados tenham que ser disponibilizados publicamente.[1] No entanto, a "Lei de Liberdade de Informação" (*Freedom of Information Act* – 5 U.S.C. § 552) garante que qualquer pessoa nos Es-

[1] Esta política é parte de um acordo de cavalheiros entre as agências financiadoras e a comunidade científica. Esse acordo funciona porque a comunidade científica até agora tem demonstrado a sua boa vontade em tornar os dados disponíveis gratuitamente e em tempo hábil. Em geral, ecólogos e cientistas da área ambiental demoram mais para disponibilizar seus dados que os cientistas que trabalham em áreas mais orientadas para o mercado (como a biotecnologia e a genética molecular), onde há maior interesse público. Da mesma forma, a preparação para o acesso aos dados e as normas para a produção dos metadados por parte dos ecólogos e cientistas ambientais têm ficado aquém das outras áreas do conhecimento. Pode-se argumentar que a falta de um GenBank para os ecólogos tem atrasado o progresso da Ecologia e das Ciências Ambientais. No entanto, dados ecológicos são muito mais heterogêneos que as sequências gênicas. Por isso, delinear uma única estrutura que facilite e acelere o compartilhamento de dados têm se mostrado uma tarefa difícil.

tados Unidos pode solicitar aos cientistas, a qualquer momento, dados coletados com recursos públicos. Por todas estas razões, encorajamos você a gerenciar seus dados com o mesmo cuidado que os coleta, analisa e publica.

O PRIMEIRO PASSO: GESTÃO DE DADOS BRUTOS

Nas ciências de bancada, o caderno de notas do laboratório é o meio tradicional para registrar as observações. Em contrapartida, dados ecológicos e ambientais são registrados de diversas formas. Exemplos incluem observações escritas em cadernos, rabiscadas em placas de plástico de mergulho ou em papel a prova d'água, ou ditada em gravadores de voz. As medidas de instrumentos digitais são transmitidas diretamente para registradores de dados (*data loggers*), computadores ou *palmtops*, e negativos de óxido de prata ou imagens digitais de satélite. Portanto, a primeira responsabilidade com nossos dados é transferi-los rapidamente de suas origens heterogêneas para um formato comum que pode ser organizado, conferido, analisado e compartilhado.

Planilhas

No Capítulo 5, ilustramos a forma de organizar os dados em uma **planilha**. Trata-se de uma página eletrônica[2] onde cada linha representa uma única observação (p. ex., a unidade de estudo; como uma folha ou pena, um organismo, ou uma ilha), e cada coluna representa uma única medida ou variável observada. A entrada em cada **célula** da planilha, ou no par de linha-por-coluna, é o valor medido para a unidade de estudo (a linha) ou da variável observada (coluna)[3]. Todos os tipos de dados utilizados por ecólogos e cientistas ambientais podem ser armazenados em planilhas e a maioria dos delineamentos experimentais (Capítulo 7) pode ser representada neste formato.

Após a coleta, os dados de campo devem ser inseridos em planilhas assim que possível. Em nossas pesquisas, tentamos transferir os dados de nossas notas de campo ou dos instrumentos para planilhas no mesmo dia. Há várias razões

[2] Embora os programas comerciais de planilhas eletrônicas possam facilitar a organização e gestão de dados, não é apropriado usá-los para o arquivamento eletrônico dos seus dados. Em vez disso, os dados arquivados devem ser armazenados como arquivos de texto ASCII (*American Standard Code for Information Exchange*), que podem ser lidos por qualquer máquina sem recorrer a programas comerciais que podem não existir mais daqui a 50 anos.

[3] Ilustramos e trabalhamos apenas com planilhas de duas dimensões (linha e coluna, também chamadas de arquivos simples). Independentemente de os dados serem armazenados em formatos com duas ou três dimensões, páginas dentro de arquivos eletrônicos devem ser exportadas como arquivos simples para serem analisadas em programas estatísticos. Embora os programas de planilhas frequentemente contenham rotinas estatísticas e geradores de números aleatórios, não recomendamos o uso desses programas para análise de dados científicos. Planilhas eletrônicas são adequadas para gerar estatísticas descritivas simples, mas os seus geradores de números aleatórios e algoritmos estatísticos podem não ser confiáveis para análises mais complexas (Knüsel, 1998; McCullough & Wilson, 1999).

importantes para transcrever seus dados imediatamente. Primeiro, você pode verificar, com rapidez, se observações ou unidades experimentais necessárias foram perdidas devido ao descuido do pesquisador ou à falha de instrumento. Se for necessário voltar e coletar observações adicionais, isto deve ser feito rapidamente ou os novos dados não serão comparáveis. Segundo, suas folhas de anotações normalmente incluirão observações e anotações nas margens, além dos números registrados. Essas notas nas margens fazem sentido logo que você as faz, porém, elas têm uma vida curta e rapidamente tornam-se incompreensíveis (algumas vezes, até ilegíveis). Terceiro, uma vez que os dados são inseridos em uma planilha, fique com duas cópias, é menos provável você extraviar ou perder ambas as cópias de seus valiosos resultados. Por fim, a rápida organização dos dados pode agilizar a análise e a publicação; os dados brutos desorganizados certamente não.

Os dados devem ser conferidos o mais rápido possível após serem digitados em uma planilha. No entanto, sua revisão geralmente não detecta todos os erros e, uma vez que os dados tenham sido organizados e catalogados, verificações adicionais são necessárias (assunto discutido posteriormente, neste capítulo).

Metadados

Ao mesmo tempo que você insere os dados em uma planilha, deve começar a construir os **metadados** (que devem acompanhar os dados). São "dados sobre dados" e descrevem os atributos-chave do conjunto de dados. Os metadados[4] que acompanham um conjunto de dados devem incluir, no mínimo:

- o nome e as informações de contato de quem coletou os dados;
- a informação geográfica sobre o local onde os dados foram coletados;
- o nome do estudo em que os dados foram coletados;
- as fontes de apoio que permitiram sua coleta;
- uma descrição da organização do arquivo de dados, que deverá incluir:
  - uma breve descrição dos métodos usados para coletar os dados;
  - os tipos de unidades experimentais;
  - as unidades de medida ou observação de cada uma das variáveis;
  - a descrição de quaisquer abreviações utilizadas no arquivo de dados;

[4] Três "níveis" de metadados foram descritos por Michener et al. (1997) e Michener (2000). Os metadados de nível I incluem uma descrição básica do conjunto de dados e da estrutura dos dados. Metadados de nível II contêm não apenas as descrições e a estrutura dos conjuntos de dados, mas também informações sobre a origem e a localização da pesquisa, informações sobre a versão dos dados e instruções sobre como acessar os dados. Por fim, os metadados de nível III são considerados passíveis de auditoria e publicáveis. Os metadados de nível III incluem todas as informações daqueles de nível II mais descrições de como foram obtidos, da documentação do processo de garantia de qualidade e do controle de qualidade (GQ/CQ), uma descrição completa de qualquer programa utilizado no processamento dos dados, uma descrição clara de como os dados foram arquivados, um conjunto de publicações relacionadas aos dados e uma história de como o conjunto de dados foi utilizado.

- uma descrição explícita de quais dados estão nas colunas, quais estão nas linhas e qual caractere foi usado para separar elementos dentro das linhas (p. ex., tabulações, espaços ou vírgulas) e entre colunas (p. ex., um "enter").[5]

Uma abordagem alternativa, mais comum nas ciências de bancada, é criar os metadados *antes* de os dados serem coletados. Uma boa parte do raciocínio claro pode acompanhar a exposição formal e a redação dos métodos, incluindo o delineamento amostral, o número de observações previstas, o espaço amostral, e a organização do arquivo de dados. Tal construção *a priori* dos metadados também facilita a redação da seção de métodos do relatório final ou da publicação do estudo. Também apoiamos que gráficos e tabelas sejam traçados para ilustrar as relações entre as variáveis que se pretende examinar.

Independente de você optar por escrever seus metadados antes ou após a realização do seu estudo, não podemos exagerar sua importância. Mesmo pensando que escrever os metadados consuma tempo e que eles pareçam conter informações evidentes, garantimos que você irá considerá-los de valor inestimável. Você pode achar que nunca esquecerá que "Charco Nemo" foi onde coletou os dados de sua obra-prima sobre orquídeas ameaçadas de extinção. E é claro que ainda lembrará de todas as estradas vicinais que percorreu até chegar lá. Mas o tempo irá corroer essas memórias conforme os meses e anos passam, e à medida que você se envolver com novos projetos e estudos.

Todos nós já temos experiências de olhar listas de nomes em arquivos de computador (NEMO, NEMO03, NEMO03NOVO) tentando encontrar o que precisamos. Mesmo se pudermos encontrar o arquivo, ainda podemos estar sem sorte. Sem metadados, não há nenhuma maneira de recordar o significado da sopa de letrinhas das abreviaturas e siglas na linha superior do arquivo. Reconstruir conjuntos de dados é frustrante, ineficiente, consome tempo e impede a análise mais avançada dos dados. Conjuntos de dados não documentados são praticamente inúteis para você e para qualquer pessoa. Dados sem os metadados não podem ser armazenados em arquivos públicos acessíveis e não têm vida útil além das estatísticas descritivas em uma publicação ou um relatório.

O SEGUNDO PASSO: ARMAZENAMENTO E CURADORIA DE DADOS

Armazenamento: temporário e permanente

Uma vez que os dados estão organizados e documentados, o conjunto de dados – uma combinação de dados e metadados – deve ser armazenado em um

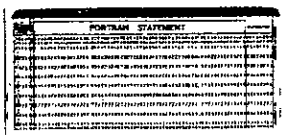
[5] Ecólogos e cientistas da área ambiental estão desenvolvendo protocolos para metadados. Esses protocolos incluem o "Conteúdo Padrão para Metadados Geoespaciais Digitais" (*Content Standards for Digital Geospatial Metadata;* FGDC, 1998) para dados espacialmente organizados (p. ex., dados de Sistemas de Informação Geográfica ou SIG), descritores-padrão para dados ecológicos e ambientais de solos (Boone et al., 1999), e de classificação e normas de informação para dados que descrevem tipos de vegetação (FGDC, 1997). Muitos desses protocolos foram incorporados na Linguagem Ecológica de Metadados (ou EML, do inglês "*Ecological Metadata Language*"), desenvolvido no final de 2002 (http://knb.ecoinformatics.org/software/eml/).

meio (ou mídia) permanente. O sistema mais durável, e o único meio aceitável como verdadeiramente permanente, é o papel livre de acidez. É uma boa prática imprimir uma cópia da planilha de dados brutos, acompanhada de seus metadados, em papel livre de acidez usando uma impressora a *laser*.[6] Essa cópia deve ser armazenada em algum lugar seguro contra danos. O armazenamento eletrônico do conjunto de dados original também é uma boa ideia, mas você não deve esperar que as mídias eletrônicas durem mais que 5 a10 anos.[7] Assim, o armazenamento eletrônico deve ser usado principalmente para cópias em uso. Se você tiver a intenção de manter cópias eletrônicas por mais alguns anos, terá que copiar os conjuntos de dados para as novas mídias eletrônicas com regularidade. É preciso reconhecer que o armazenamento de dados tem custos reais, de espaço (onde os dados são armazenados), de tempo (para manter e transferir conjuntos de dados entre as mídias) e de dinheiro (porque o tempo e o espaço valem muito).

Curadoria de dados

A maioria dos dados ecológicos e ambientais é coletada por pesquisadores usando recursos obtidos através de financiamentos e contratos. Em geral, esses dados pertencem tecnicamente à agência financiadora, e eles devem estar disponíveis a terceiros dentro de um período de tempo razoavelmente curto. Desta forma, independente de como seus conjuntos de dados são armazenados, devem ser mantidos de forma que os tornem usáveis não só por você, mas também pela ampla comunidade científica e por outros indivíduos ou grupos interessados. Como em museus e herbários ou coleções de animais e plantas, os conjuntos de dados devem ser gerenciados para que estejam acessíveis a outras pessoas. Vários conjuntos de dados gerados por um projeto ou grupo de

[6] A tinta das impressoras matriciais antigas (e máquinas de escrever) dura mais que das impressoras jato-de-tinta, mas a ligação eletrostática da tinta no papel das impressoras a *laser* tem durabilidade superior às demais. Os verdadeiros Luditas preferem a tinta Higgins® Aeterna India (em reverência a Henry Horn).



Cartão perfurado de computador

[7] Poucos leitores se lembrarão dos disquetes de qualquer tamanho (8", 5-1/4", 3-1/2"), muito menos das fitas de papel, dos cartões perfurados (ilustração) ou dos rolos de fita magnética. Enquanto produzimos este livro, os onipresentes CD-ROMs estão sendo substituídos por DVDs, e provavelmente teremos que comprar nossas coleções de discos de vinil pela segunda vez em algumas décadas. Embora a vida útil da maioria dos mídias magnéticas é da ordem de anos a décadas, e que a maioria dos meios ópticos (CDs. DVDs) é da ordem de décadas a séculos (apenas de anos nos trópicos úmidos, onde fungos comem a cola entre as camadas de material laminado e estragam o disco), o mercado capitalista os substituem em um ciclo de aproximadamente 5 anos. Assim, embora ainda tenhamos em nossas prateleiras disquetes 5-1/4" legíveis com dados dos anos 1980's e 1990's, é quase impossível encontrar em 2004, muito menos comprar, um leitor de disquetes para os ler. Mesmo se pudéssemos encontrar um, os sistemas operacionais atuais dos nossos computadores não irão reconhecê-los como um equipamento periférico. Uma boa fonte de equipamentos e programas obsoletos é o *"Computer History Museum"*, em Mountain View, Califórnia: http://www.computerhistory.org/.

pesquisa podem ser organizados em catálogos eletrônicos de dados. Esses catálogos muitas vezes podem ser acessados e pesquisados de localidades remotas usando a Internet.

No momento em que escrevemos este livro, não existem padrões estabelecidos para a curadoria de dados ecológicos e ambientais. A maioria dos catálogos de dados é mantida em servidores conectados à rede mundial de computadores (*World Wide Web*),[8] e pode ser encontrado usando as ferramentas de busca disponíveis. Assim como o armazenamento de dados, sua curadoria é cara: estima-se que a gestão e a curadoria de dados devem consumir cerca de 20% do custo total de um projeto de pesquisa (Michener & Haddad, 1992). Infelizmente, embora as agências financiadoras exijam que os dados sejam disponibilizados, elas têm sido relutantes em financiar os custos totais do gerenciamento e da curadoria de dados. Os próprios pesquisadores raras vezes orçam esses custos em seus pedidos de financiamento. Quando os orçamentos são inevitavelmente cortados, os custos de gestão e curadoria de dados em geral são os primeiros itens a ser descartados.

O TERCEIRO PASSO: VERIFICAÇÃO DOS DADOS

Antes de começar a analisar um conjunto de dados, você deve examiná-lo cuidadosamente para detectar as discrepâncias e os erros.

A importância dos dados discrepantes

Os **dados discrepantes** (*Outliers*) são valores de medições ou observações registradas que estão fora do intervalo da maior parte dos dados (Figura 8.1). Embora não exista um nível padrão para "fora do intervalo" que defina um dado discrepante, é uma prática comum considerar valores além do decil superior ou inferior (90º percentil, 10º percentil) de uma distribuição como potenciais dados, e

[8] Muitos cientistas consideram que inserir dados na Internet é um meio de arquivamento permanente de dados. Isso é ilusório. Primeiro, isso é apenas uma transferência de responsabilidade entre você e um gestor de sistemas de informação (ou outro profissional da área de tecnologia da informação). Ao colocar sua cópia de arquivos eletrônica na Internet, você espera que cópias de segurança serão feitas com regularidade e mantidas pelo gestor do sistema. Cada vez que um servidor é aprimorado, os dados têm de ser copiados do antigo servidor para o novo. A maioria dos laboratórios ou departamentos não tem seus próprios gestores de sistemas e os interesses dos centros de computação em arquivar e manter páginas e arquivos de dados na Internet não são necessariamente iguais aos dos investigadores. Segundo, os discos rígidos do servidor falham com frequência (e muitas vezes espetacularmente). Por último, a Internet não é permanente nem estável. GOPHER e LYNX desapareceram, FTP está sendo substituída por HTTP e HTML, e a linguagem atual da *web* já está sendo progressivamente abandonada em favor da (não inteiramente compatível) XML. Todas essas alterações na funcionalidade da Internet e da acessibilidade dos arquivos armazenados ocorreram em 10 anos. Muitas vezes, é mais fácil recuperar dados de cadernos que foram escritos à mão no século XIX que os de sítios da Internet, que foram digitalmente "arquivados" na década de 1990!

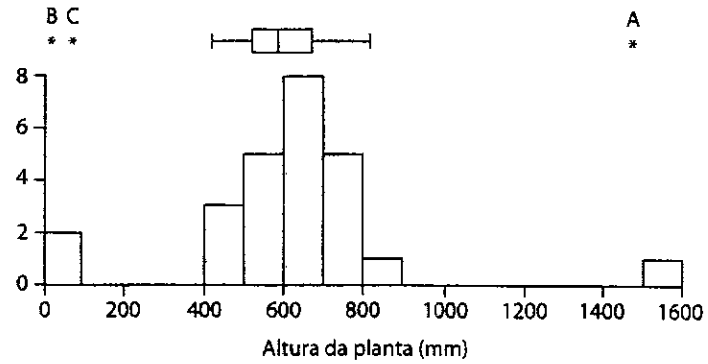


Figura 8.1 Histograma e *box plot* de amostras das medidas da altura de plantas-jarro carnívoras, *Darlingtonia californica* (tamanho amostral $N = 25$; dados da Tabela 8.1). O histograma mostra a frequência das plantas em classes de tamanho com intervalos de 100 mm. Acima do histograma está representado um *box plot* dos mesmos dados. A linha vertical dentro da caixa indica a mediana da distribuição. A caixa inclui 50% dos dados, e os "bigodes" abrangem 90%. A, B e C indicam os três dados extremos e correspondem às linhas rotuladas na Tabela 8.1. Os histogramas e *box plots* permitem que você identifique com rapidez os valores extremos de seus dados.

aqueles menores do que o 5º percentil ou superior ao 95º percentil como possíveis valores extremos. Muitos pacotes estatísticos destacam esses valores em *box plots* (como fizemos com estrelas e letras na Figura 8.1) para indicar que você deve examiná-los com cuidado. Alguns desses pacotes também possuem ferramentas interativas, descritas na próxima seção, que permitem a localização desses valores dentro da planilha.

É importante identificar dados discrepantes, pois eles podem ter efeitos dramáticos sobre seus testes estatísticos. Dados discrepantes e valores extremos aumentam a variância dos dados. Variâncias infladas diminuem o poder do teste e aumentam a possibilidade de erro Tipo II (falha em rejeitar uma hipótese nula falsa, *ver* Capítulo 4). Por essa razão, alguns pesquisadores excluem automaticamente dados discrepantes e valores extremos antes de realizar suas análises. Essa é uma prática ruim! Existem apenas duas razões justificáveis para descartar dados: (1) os dados estão errados (p. ex., eles foram anotados incorretamente no caderno de campo) e (2) os dados não representam observações válidas do seu espaço amostral original (p. ex., uma de suas parcelas em dunas foi invadida por veículos *off-road*). Excluir observações simplesmente porque elas são "confusas" é a lavagem* ou adulteração de resultados e, legitimamente, pode ser considerada fraude científica.

Os dados discrepantes são mais do que apenas um ruído. Podem refletir processos biológicos reais e uma reflexão cuidadosa sobre o seu significado, levando a

* N. de T. Do inglês *laundering*, é como "lavagem de dinheiro".

novas ideias e hipóteses.[9] Além disso, alguns dados irão parecer discrepantes apenas porque estão sendo forçados a se ajustar a uma distribuição normal. Os dados de distribuições não normais, como a log-normal ou a exponencial, muitas vezes parecem ser extremos, mas se adequam após serem devidamente transformados (*ver* abaixo).

Erros

Os **erros** são valores registrados que não representam as medidas ou observações originais. Nem todos os erros são dados discrepantes. Da mesma forma, nem todos os dados discrepantes necessariamente são erros. Os erros podem aparecer no conjunto de dados de duas maneiras: como erros de coleta ou como de transcrição. No primeiro caso, são dados que resultam de instrumentos quebrados ou descalibrados, ou da anotação errada de dados no campo. Além disso, podem ser difíceis de identificar. Se você ou seu assistente de campo estiverem anotando os dados e escreverem um valor incorreto, isto raramente poderá ser corrigido.[10] A menos que o erro seja reconhecido imediatamente durante a transcrição dos dados, caso contrário ele provavelmente não será detectado e contribuirá para a variação dos dados.

Os erros resultantes de instrumentos quebrados ou descalibrados são mais fáceis de detectar (determina-se que o instrumento está quebrado ou após ele ser

[9] Embora os biólogos sejam treinados para pensar em medianas e médias, e usar a estatística para testar padrões de tendência central dos dados, informações ecológicas e evolutivas interessantes acontecem muitas vezes na cauda de uma distribuição estatística. De fato, existe uma parte da estatística dedicada ao estudo e à análise de valores extremos (Gaines & Denny, 1993), que podem ter impactos duradouros em ecossistemas (p. ex., Foster et al., 1998). Por exemplo, populações isoladas (*sensu* Mayr, 1963), que resultam de um único evento fundador, podem ser suficientemente distintas em sua composição genética, pois seriam estatisticamente consideradas dados discrepantes em relação à população original. Tais populações isoladas, se sujeitas a novas pressões seletivas, poderiam se tornar reprodutivamente isoladas de sua população original e formar novas espécies (Schluter, 1996). As reconstruções filogenéticas sugerem que populações isoladas podem resultar em novas espécies (p. ex., Green et al., 2002), mas testes experimentais da teoria não apoiaram o modelo de especiação pelo efeito do fundador e populações isoladas (Mooers et al., 1999).

[10] Para minimizar erros de coleta no campo, muitas vezes nos envolvemos em um tipo de diálogo de perguntas e respostas entre nós e nossos assistentes de campo:

Coletor de dados: "Esta é a planta 107. Ela tem 4 folhas, não tem filódios, não tem flores."
Anotador dos dados: "Certo, 4 folhas, sem filódios e sem flores para a 107."
Coletor de dados: "Onde está a 108?"
Anotador dos dados: "Morreu no ano passado. A 109 deve estar cerca de 10 metros a sua esquerda. No ano passado ela tinha 7 folhas e uma flor atrofiada. Depois da 109, deve ter apenas mais uma planta nesta parcela."

Repetir as observações passadas de um para o outro e manter um controle cuidadoso das informações ajuda a minimizar os erros de coleta. Esta "conversa sobre dados" também nos mantêm atentos e concentrados durante longos dias de campo.

recalibrado). Infelizmente, o erro do instrumento pode resultar na perda de uma grande quantidade de dados. Experimentos que dependem da coleta automatizada de dados necessitam de procedimentos adicionais para checagem e manutenção de equipamentos, minimizando, com isso, a perda de dados. Pode ser possível coletar valores de reposição para dados que resultam de falhas de equipamento, mas apenas se as falhas forem detectadas logo após a coleta. As possíveis falhas de equipamentos são outra razão para transcrever logo seus dados brutos para as planilhas e avaliar a sua acurácia.

Os erros de transcrição resultam principalmente de má digitação dos valores nas planilhas. Quando esses erros parecerem discrepâncias, podem ser conferidos com as fichas de campo originais. Provavelmente, o erro de transcrição mais comum é colocar a vírgula decimal fora de lugar, alterando a ordem de grandeza do valor. Quando os erros de transcrição não aparecem como dados discrepantes, podem permanecer não detectados, a menos que as planilhas sejam conferidas com cuidado. Os erros na transcrição são menos comuns quando os dados dos instrumentos são transferidos eletronicamente direto para as planilhas. No entanto, podem ocorrer erros de transmissão, que podem resultar em erros de "deslocamento de estrutura", nos quais um valor colocado na coluna errada resulta no deslocamento de todos os valores subsequentes para colunas incorretas. A maioria dos instrumentos possui programas para checar erros de transmissão e relatá-los ao usuário em tempo real. Os arquivos de dados originais procedentes de instrumentos de coleta automatizada de dados nunca devem ser apagados da memória do instrumento até que as planilhas tenham sido conferidas com cautela.

Dados faltantes

Um problema relacionado, e igualmente importante, é o tratamento de dados faltantes (*missing values*). Seja muito cuidadoso em suas planilhas e em seus metadados para distinguir os valores medidos como 0 dos dados faltantes (réplicas em que nenhuma observação foi registrada). Os dados faltantes em uma planilha devem receber uma designação própria (como a sigla "ND*" para "não disponível") em vez de deixar as células em branco. Isso merece cuidado, pois alguns programas podem tratar as células em branco como zeros (ou inserir "0" nos espaços em branco), o que irá arruinar a sua análise.

Detectando discrepâncias e erros

Encontramos três técnicas especialmente úteis para detectar dados discrepantes e erros em conjuntos de dados: calcular estatísticas de colunas, checar os intervalos e a precisão dos valores das colunas, e analisar dados de maneira exploratória com gráficos. Outros métodos mais complexos foram discutidos por Edwards (2000). Usamos como exemplo para detecção de discrepâncias a medida das

* N. de T. A sigla depende do programa que for utilizar e, em geral, emprega-se a em inglês (NA para *not available*).

alturas das folhas-jarro da *Darlingtonia californica* (lírio-cobra, ou planta-jarro da Califórnia).[11] Coletamos esses dados, apresentados na Tabela 8.1, como parte de um estudo de crescimento, alometria e fotossíntese dessa espécie. Neste exemplo com apenas 25 observações, os dados discrepantes são fáceis de detectar, simplesmente olhando as colunas de números. Entretanto, a maioria dos conjuntos de dados ecológicos terá muito mais observações e será mais difícil encontrar valores incomuns no arquivo de dados.

ESTATÍSTICAS DE COLUNA O cálculo de estatísticas de coluna simples dentro da planilha é uma maneira fácil de identificar valores inusitadamente grandes ou pequenos no conjunto de dados. As medidas de posição e dispersão, como média, mediana, desvio-padrão e variância das colunas, dão uma visão geral e rápida da distribuição dos valores nas colunas. Os valores mínimo e máximo em uma coluna podem indicar resultados grandes ou pequenos. A maioria dos programas de planilhas eletrônicas possui funções que calculam esses valores. Se inserir essas funções nas seis últimas linhas da planilha (Tabela 8.1), você poderá usá-las como uma primeira checagem dos dados. Se encontrar um valor extremo e determinar que é um erro verdadeiro, a planilha atualizará automaticamente os cálculos quando você substituir o valor pelo número correto.

CHECANDO O INTERVALO E A PRECISÃO DOS VALORES DAS COLUNAS Outra forma de verificar os dados é usar as funções das planilhas para assegurar que os valores em uma determinada coluna estão dentro de limites razoáveis ou se refletem a precisão das medidas. Por exemplo, os jarros de *Darlingtonia* raras vezes ultrapassam os 900 mm de altura. Qualquer medida muito acima desse valor seria suspeita. Poderíamos olhar com atenção o conjunto de dados para ver se quaisquer valores são extremamente grandes, porém, para conjuntos de dados reais, com centenas ou milhares de observações, a verificação manual é ineficiente e sem acurácia. A maioria dos programas de planilhas eletrônicas tem funções lógicas que podem ser utilizadas para checagem de valores e automatização do processo. Desse modo, você poderia usar a declaração:

Se (valor a ser conferido) > (um valor máximo), registre um "1"; caso contrário, registre "0" para verificar rapidamente se qualquer valor de altura excede um limite superior (ou inferior).

Darlingtonia californica

[11] A *Darlingtonia californica* é uma planta carnívora da família Sarraceniaceae que cresce apenas nos charcos (*"serpentine fens"*) das montanhas Siskiyou em Oregon e Califórnia e nas montanhas de Sierra Nevada na Califórnia. Suas folhas em forma de jarro normalmente atingem 800 mm de altura, mas às vezes passam de 1 m. Essas plantas alimentam principalmente de formigas, moscas e vespas.

TABELA 8.1 Conjunto de dados ecológicos ilustrando valores extremos típicos para *Darlingtonia californica*

	Planta #	Altura	Boca	Tubo
	1	744	34,3	18,6
	2	700	34,4	20,9
	3	714	28,9	19,7
	4	667	32,4	19,5
	5	600	29,1	17,5
	6	777	33,4	21,1
	7	640	34,5	18,6
	8	440	29,4	18,4
	9	715	39,5	19,7
	10	573	33,0	15,8
A	11	1.500	33,8	19,1
	12	650	36,3	20,2
	13	480	27,0	18,1
	14	545	30,3	17,3
	15	845	37,3	19,3
	16	560	42,1	14,6
	17	450	31,2	20,6
	18	600	34,6	17,1
	19	607	33,5	14,8
	20	675	31,4	16,3
	21	550	29,4	17,6
B	22	5,1	0,3	0,1
	23	534	30,2	16,5
	24	655	35,8	15,7
C	25	65,5	3,52	1,77
Média		611,7	30,6	16,8
Mediana		607	33,0	18,1
Desvio-Padrão		265,1	9,3	5,1
Variância		70.271	86,8	26,1
Mínimo		5,1	0,3	0,1
Máximo		1.500	42,1	21,1

Os dados são medidas morfológicas de indivíduos de plantas-jarro. A altura do jarro (em mm), o diâmetro da boca do jarro (em mm) e o diâmetro do tubo do jarro (em mm) foram registrados para 25 indivíduos amostrados em Days Gulch. As linhas com valores extremos na Figura 8.1 estão realçadas em negrito e designadas como A, B e C. Estatísticas descritivas para cada variável estão na parte inferior da tabela. É crucial identificar e avaliar os valores extremos em um conjunto de dados. Tais valores aumentam muito a estimativa da variância, e é importante averiguar se essas observações representam erros de medidas, medidas atípicas ou apenas a variação natural da amostra.

Este mesmo método pode ser usado para conferir a precisão dos valores. Medimos a altura utilizando uma fita métrica com marcações em milímetros, e, por costume, arredondamos para o milímetro mais próximo. Assim, nenhum valor na planilha deve ser mais preciso que um milímetro (i. e., nenhum valor deve ter frações de milímetros). A função lógica para checar isso é:

> Se (o valor a ser conferido tiver qualquer outro que não um 0 depois da vírgula decimal), registre um "1"; caso contrário, registre "0".

Você pode imaginar diversas formas similares para checar os valores em colunas, e a maioria pode ser automatizada, utilizando funções lógicas dentro de planilhas eletrônicas.

A checagem do intervalo dos dados da Tabela 8.1 revela que a altura registrada para a planta 11 (1.500 mm) é suspeita, pois essa observação excede o limite de 900 mm. Além disso, as alturas registradas para a planta 22 (5,1 mm) e para a planta 25 (65,5 mm) foram digitadas erroneamente, pois os valores decimais são maiores que a precisão das nossas fitas métricas. Esta auditoria poderia nos levar a re-examinar nossos dados nas fichas de campo ou, talvez, voltar ao campo para medir as plantas novamente.

ANÁLISE DE DADOS GRÁFICA EXPLORATÓRIA As figuras científicas – gráficos – são uma das ferramentas mais poderosas que você pode usar para descrever os resultados de seu estudo (Tufte, 1986, 1990; Cleveland, 1985, 1993). Eles também podem ser usados para detectar erros e discrepâncias, e para evidenciar padrões inesperados nos seus dados. A utilização de gráficos antes da análise formal dos dados é chamada de **análise de dados gráfica exploratória** ou gráficos EDA*, ou simplesmente EDA (p. ex., Tukey, 1977; Ellison, 2001). Aqui, concentramo-nos no uso da EDA para a detecção de dados discrepantes e valores extremos. O uso da EDA para procurar por padrões, novos ou inesperados, nos dados ("mineração de dados" ou, mais pejorativamente, "garimpagem ou dragagem de dados") está evoluindo em sua própria indústria artesanal (Smith & Ebrahim, 2002).[12]

[12] A objeção estatística à dragagem de dados é que se fizermos bastante gráficos a partir de um conjunto de dados grande e complexo, certamente iremos encontrar algumas relações que pareçam estatisticamente significativas. Assim, a dragagem de dados debilita nosso cálculo dos valores de *P* e aumenta as chances de rejeitarmos incorretamente a hipótese nula (erro Tipo I). A objeção filosófica à EDA é que você já deve saber de antemão como seus dados serão plotados e analisados. Como enfatizamos nos Capítulos 6 e 7, a definição prévia do delineamento amostral e das análises caminha lado a lado com uma questão de pesquisa bem focada. Por outro lado, a EDA é essencial para a detecção de valores discrepantes e erros. Não há como negar que você pode revelar padrões interessantes apenas casualmente, plotando seus dados. Você sempre pode voltar ao campo para testar esses padrões com um delineamento experimental apropriado. Bons programas estatísticos facilitam análises EDA, porque os gráficos podem ser construídos com rapidez e facilmente visualizados na tela do computador.

* N. de T. EDA é a abreviação, em inglês, para *Exploratory Data Analysis* (análise de dados exploratória).

```
    0   06
    1
    2
    3
    4   458
D   5   34567
M   6   00145577
D   7   01148
    8   4
    9
   10
   11
   12
   13
   14
   15   0
```

Figura 8.2 Gráfico de caule-e-folha. Os dados são medidas da altura (em mm) de 25 indivíduos de *Darlingtonia californica* (Tabela 8.1). Neste gráfico, a primeira coluna de números à esquerda – o "caule" — representa a casa das "centenas" e cada "folha", à direita, a casa das "dezenas" de cada valor registrado. Por exemplo, a primeira linha ilustra as observações 00x e 06x, ou uma observação menor que 10 (o valor de 5,1 da planta 22, na Tabela 8.1), e outra observação entre 60 e 70 (o valor de 65,5 da planta 25). A quinta linha ilustra as observações 44x, 45x e 48x, que correspondem a valores de 440 (planta 48), 450 (planta 17) e 480 (planta 13). A última linha ilustra a observação 150x, correspondente ao valor 1.500 (planta 11). Todas as observações contidas na coluna 1 da Tabela 8.1 estão incluídas nessa representação. A mediana (607) está na linha 7, indicada com a letra M. O quartil inferior (545), também conhecido como **dobradiça**, está na linha 6, indicado com a letra D. O quartil superior ou dobradiça (700) está na linha 8, também marcado com a letra D. Da mesma forma que histogramas e *box plots*, gráficos de caule-e-folha são uma ferramenta de diagnóstico para visualizar seus dados e detectar dados extremos.

Existem três tipos de gráficos indispensáveis para detectar dados discrepantes e valores extremos: ***box plots*, gráficos de caule-e-folha** e **gráficos de dispersão**. Os dois primeiros são melhores para dados univariados (plotar a distribuição de uma única variável), enquanto os gráficos de dispersão são utilizados para dados bivariados ou multivariados (plotar as relações entre duas ou mais variáveis). Você já viu muitos exemplos de dados univariados neste livro: as 50 observações do comprimento do espinho tibial de aranhas (*ver* Figura 2.6 e Tabela 3.1), a presença ou ausência de *Rhexia* em municípios de Massachusetts (Capítulo 2), e a taxa reprodutiva do besouro girinídeo de manchas na cor laranja (Capítulo 1) são todos os conjuntos de medidas de uma única variável. Para ilustrar a detecção de dados discrepantes e gráficos univariados de EDA, usamos a primeira coluna de dados da Tabela 8.1 – altura do jarro.

Nossas estatísticas descritivas na Tabela 8.1 sugerem que podem haver alguns problemas com os dados. A variância é grande em relação à média, e tanto o valor mínimo quanto o máximo parecem extremos. Um *box plot* (no topo da Figura 8.1) mostra que dois pontos são inusitadamente pequenos e um ponto é bastante grande. Podemos identificar esses valores inusitados em mais detalhes com um gráfico de caule-e-folha (Figura 8.2), que é uma variante do familiar histograma

(Capítulo 1; Tukey, 1977). O gráfico de caule-e-folha mostra claramente que os dois valores pequenos são de uma ordem de grandeza (i. e., cerca de dez vezes) menor que a média ou que a mediana, e que o valor grande é mais do que o dobro da média ou da mediana.

Quão inusitados são esses valores? Sabe-se que alguns jarros de *Darlingtonia* excedem um metro de altura, e jarros muito pequenos também podem ser encontrados. Talvez essas três observações simplesmente reflitam a variabilidade inerente ao tamanho dos jarros do lírio-cobra. Sem informações adicionais, não há razão para pensar que estão errados, ou que não sejam parte do nosso espaço amostral, e seria inapropriado removê-los do conjunto de dados antes das análises estatísticas posteriores.

No entanto, ainda temos informações adicionais. A Tabela 8.1 mostra outras duas medidas que fizemos a partir dos mesmos jarros: o diâmetro do tubo do jarro e o da abertura da boca do jarro. Poderíamos esperar que todas essas variáveis tivessem relação umas com as outras, tanto por causa de pesquisas anteriores sobre *Darlingtonia* (Franck, 1976) quanto porque o crescimento de partes das plantas (e animais) tende a ser correlacionado (Niklas, 1994).[13] Uma forma rápida de explorar como essas variáveis covariam umas com as outras é plotar todas as relações entre os pares possíveis (Figura 8.3).

A **matriz de gráficos de dispersão** mostrada na Figura 8.3 fornece informação adicional com as quais decide se nossos valores inusitados são discrepantes. Os dois gráficos de dispersão na linha superior da Figura 8.3 sugerem que a planta inusitadamente alta na Tabela 8.1 (ponto A; planta 11) também é alta em relação aos seus diâmetros do tubo e da boca do jarro. Porém, os

[13] Imagine um gráfico no qual você plotou o logaritmo do comprimento do jarro no eixo-x e o do diâmetro do tubo no eixo-y. Esse tipo de relação alométrica revela o modo com que a forma de um organismo pode mudar com seu tamanho. Suponha que lírios-cobra pequenos são réplicas perfeitas em escala reduzida de lírios-cobra grandes. Neste caso, o aumento no diâmetro do tubo é sincrônico com o aumento do comprimento jarro, e a inclinação da reta (β_1) = 1 (*ver* Capítulo 9). O crescimento é isométrico e as plantas pequenas parecem miniaturas das grandes. Por outro lado, suponha que jarros pequenos tenham diâmetros do tubo que são relativa, mas não absolutamente maiores que os do tubo de plantas maiores. Conforme as plantas crescem em tamanho, os diâmetros do tubo crescem relativamente menos em comparação ao comprimento dos jarros. Esse seria um caso de alometria negativa e resultaria em um valor de inclinação $\beta_1 < 1$. Por fim, se há alometria positiva, o diâmetro do tubo aumenta mais rápido com o aumento do comprimento do jarro ($\beta_1 > 1$). Esses padrões foram ilustrados por Franck (1976). Um exemplo mais conhecido é o do bebê humano, que nasce com uma cabeça relativamente grande, mas seu corpo cresce aos poucos e depois apresenta alometria negativa. Outras partes do corpo, como os dedos dos pés e das mãos, apresentam alometria positiva e crescem um pouco mais rápido.

O crescimento alométrico (positivo ou negativo) é a regra na natureza. Poucos organismos crescem isometricamente com jovens parecendo réplicas em miniatura dos adultos. Organismos mudam de forma à medida que crescem, em parte por limitações fisiológicas básicas do metabolismo e da absorção de materiais. Imagine que (simplificadamente) a forma de um organismo seja um cubo com um lado de comprimento L. O problema é que, conforme os organismos aumentam de tamanho, o volume aumenta em uma função cúbica (L^3) do comprimento (L), mas a área da superfície aumenta apenas como uma função quadrada (L^2). As demandas metabólicas

(*Continua*)

Diâmetro do tubo (mm) Diâmetro da boca (mm) Altura do jarro (mm)

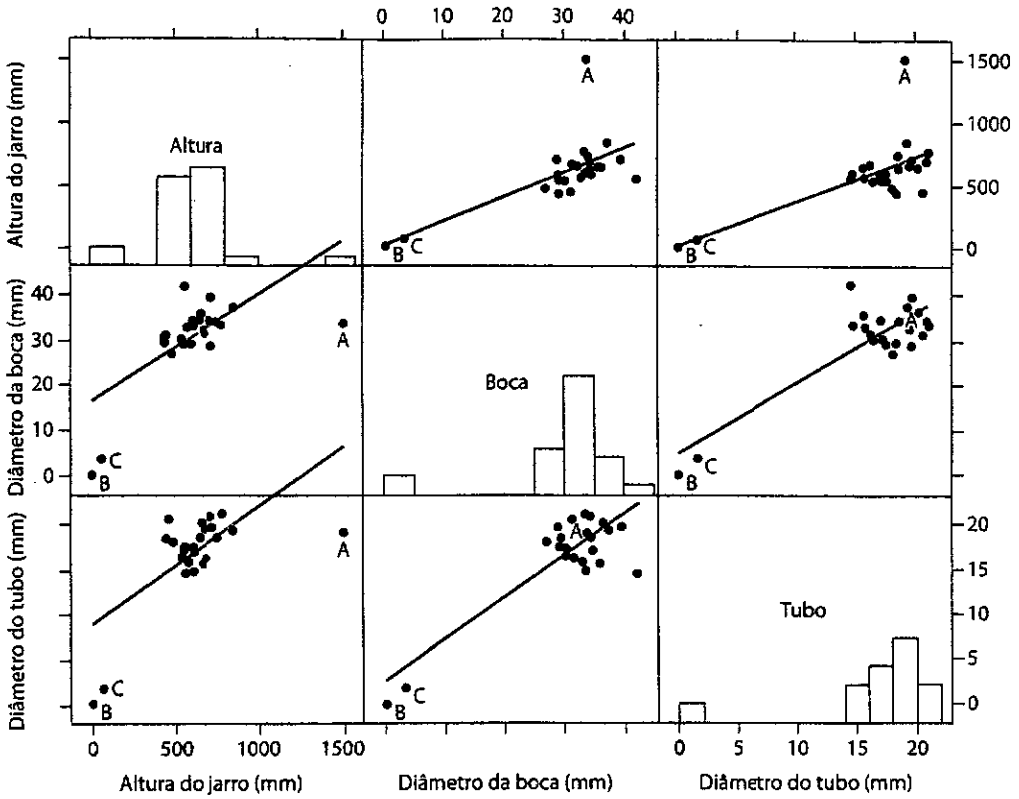


Figura 8.3 Matriz de gráficos de dispersão ilustrando a relação entre altura, diâmetro da boca e diâmetro do tubo de 25 indivíduos de *D. californica* (Tabela 8.1). Os valores extremos rotulados na Figura 8.1 e na Tabela 8.1 também estão rotulados aqui. Os painéis da diagonal nesta figura ilustram os histogramas para cada uma das 3 variáveis morfológicas. Nos painéis fora da diagonal, os gráficos de dispersão ilustram a relação entre as duas variáveis indicadas na diagonal; os gráficos de dispersão acima da diagonal são imagens-espelho daquelas abaixo da diagonal. Por exemplo, o gráfico de dispersão no canto inferior esquerdo ilustra a relação entre a altura do jarro (no eixo-*x*) e o diâmetro do tubo (no eixo-*y*), enquanto o de dispersão, no canto superior direito, mostra a relação entre o diâmetro do tubo (no eixo-*x*) e a altura do jarro (no eixo-*y*). As linhas em cada gráfico de dispersão representam o melhor ajuste da regressão linear entre as duas variáveis (*ver* Capítulo 9 para mais detalhes sobre regressão linear).

[13] (*Continuação*)

de um organismo são proporcionais ao volume (L^3), mas a transferência de material (oxigênio, nutrientes, produtos residuais) é proporcional à superfície (L^2). Portanto, as estruturas que funcionam bem para a transferência de material em pequenos organismos não funcionarão tão bem em organismos maiores. Isto fica especialmente evidente quando comparamos espécies com tamanho do corpo muito diferente. Por exemplo, a membrana da célula de um microrganismo faz um bom trabalho transferindo oxigênio por difusão simples, mas um rato ou um ser humano precisam de um pulmão vascularizado, um sistema circulatório e uma molécula de transporte especializada (hemoglobina) para realizar a mesma tarefa. Dentro e entre espécies, os padrões de forma, tamanho e morfologia muitas vezes refletem a necessidade de aumentar a área da superfície para atender às demandas fisiológicas do aumento do volume. *Ver* Gould (1977) para uma discussão detalhada sobre o significado evolutivo do crescimento alométrico.

gráficos de diâmetro da boca *versus* o diâmetro do tubo do jarro mostram que nenhuma dessas medidas são inusitadas para a planta 11. Coletivamente, esses resultados sugerem um erro – ou ao registrar a altura da planta no campo ou na transcrição de sua altura – a partir da folha de dados de campo para a planilha de computador.

Em contraste, as plantas pequenas não são inusitadas em nenhum aspecto, exceto por serem pequenas. Elas não só são pequenas como também têm tubos e bocas diminutas. Se removermos a planta 11 (a inusitadamente alta) do conjunto de dados (Figura 8.4), as relações entre todas as três variáveis das plantas pequenas não serão inusitadas em relação ao resto da população, e não haverá nenhuma razão com base nos próprios dados para suspeitar que os valores dessas plantas

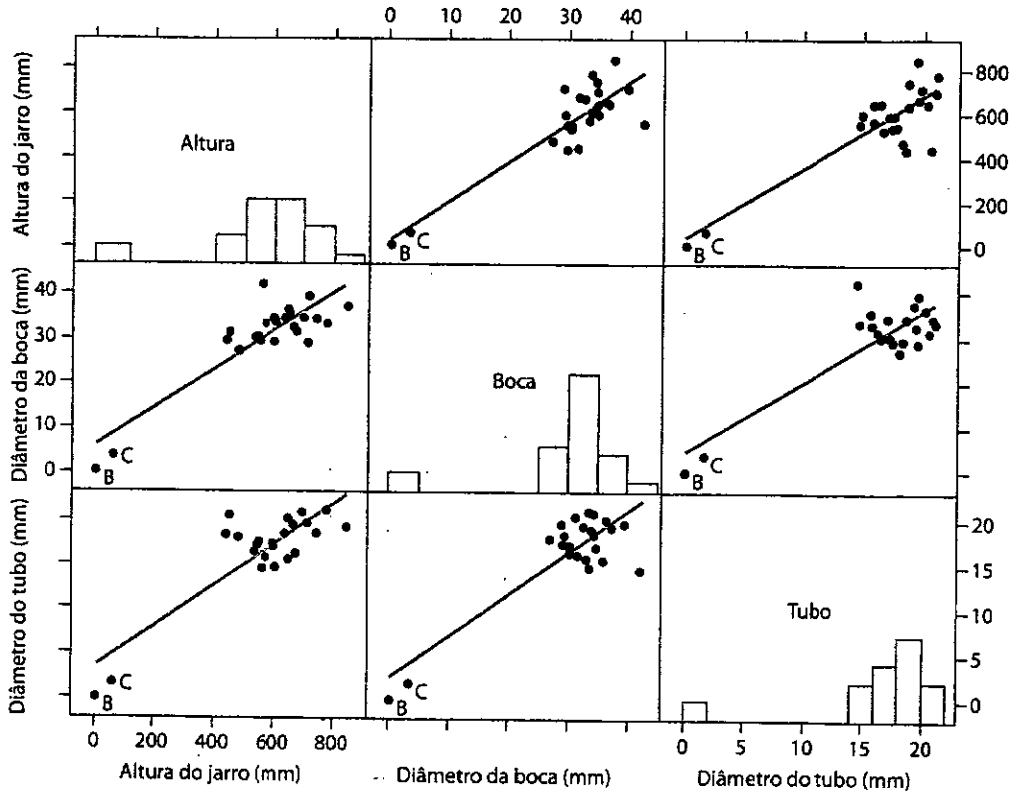


Figura 8.4 Matriz de gráficos de dispersão ilustrando a relação entre altura, diâmetro da boca e diâmetro do tubo. Os dados e a organização gráfica é a mesma da Figura 8.3. No entanto, o ponto A discrepante (Tabela 8.1) foi eliminado, o que muda algumas dessas relações. Como esse ponto era uma grande diferença na altura dos jarros (altura = 1500 mm), a escala dessa variável (eixo-*x* na coluna da esquerda) agora varia apenas de 0 a 800 mm, em comparação com 0 a 1500 mm na Figura 8.3. Não é incomum encontrar mudanças em correlações e valores de significância estatística com base na inclusão ou exclusão de um único dado com valor extremo.

foram digitados incorretamente. No entanto, essas duas plantas são muito pequenas (realmente poderíamos ter medido uma planta com 5 milímetros de altura?), motivo pelo qual devemos conferir as nossas fichas de campo e, talvez, voltar ao campo para remedir essas plantas.

Criando um rastro de auditoria

O processo de exame de um conjunto de dados para encontrar erros e dados discrepantes é um caso especial de um processo genérico de **garantia de qualidade e controle de qualidade** (abreviado como **GQ/CQ**). Como em todas as operações, em uma planilha ou em conjunto de dados, os métodos utilizados para detecção de erros e dados discrepantes devem ser documentados em uma seção de GQ/CQ nos metadados. Se for necessário remover dados da planilha, uma descrição dos valores removidos deve ser incluída nos metadados e uma nova planilha deve ser criada, contendo os dados corrigidos. Essa planilha também deve ser armazenada em mídia permanente com uma identificação exclusiva no catálogo de dados.

O processo de submeter dados a uma GQ/CQ e de modificar os arquivos de dados leva à criação de um **rastro de auditoria** dos dados. Trata-se de uma série de arquivos e de sua documentação associada que permitem a outros usuários recriar o processo pelo qual o conjunto de dados final foi analisado. O rastro de auditoria é uma parte necessária dos metadados em qualquer estudo. Se você está preparando um relatório de impacto ambiental ou outro documento utilizado em processos legais, o rastro de auditoria dos dados pode ser parte das evidências legais que apoiam o caso.[14]

O PASSO FINAL: TRANSFORMAÇÃO DE DADOS

Muitas vezes você irá ler em um artigo científico que os dados foram "transformados antes das análises". Após a transformação, os dados "atenderam aos pressupostos" do teste estatístico utilizado e a análise foi executada. Em geral, pouca coisa é dita sobre as transformações de dados, e você pode perguntar-se porquê os dados foram transformados.

Mas, primeiro, o que é uma transformação? Uma transformação significa simplesmente que uma função matemática foi aplicada a todas as observações de uma dada variável:

[14] Os historiadores da ciência têm reconstruído o desenvolvimento de teorias científicas, pelo estudo cuidadoso de rastros de auditoria negligenciados – anotações marginais rabiscadas de sucessivos rascunhos de manuscritos e em diferentes edições de monografias. O uso de processadores de texto e planilhas eliminou as anotações marginais e os grifos, mas permitiu a criação e a manutenção de rastros de auditoria formais. Outros usuários dos seus dados e futuros historiadores da ciência irão agradecer por você manter um rastro de auditoria em seus dados.

$$Y^* = f(Y) \tag{8.1}$$

O Y representa a variável original; o Y^*, a variável transformada, e f é uma função matemática aplicada aos dados. A maioria das transformações corresponde a funções algébricas simples, sujeitas simplesmente à exigência de que sejam **funções monótonas contínuas.**[15] Por serem monótonas, as transformações não alteram a ordenação dos dados, mas mudam o espaçamento relativo entre os pontos de dados e, portanto, afetam a variância e a forma da distribuição de probabilidades (Capítulo 2).

Há duas razões legítimas para transformar seus dados antes das análises. Primeiro, as transformações podem ser úteis (embora não estritamente necessárias), pois os padrões nos dados transformados podem ser mais fáceis de compreender e comunicar do que os nos dados brutos. Segundo, transformações podem ser necessárias para que a análise seja válida – este é o "atenderam aos pressupostos" usado com maior frequência em artigos científicos e discutidos nos livros de biometria.

Transformações de dados como ferramenta cognitiva

As transformações muitas vezes são úteis para converter curvas em linhas retas. Conceitualmente, relações lineares são mais fáceis de entender e muitas vezes têm melhores propriedades estatísticas (*ver* Capítulo 9). Quando duas variáveis são relacionadas uma com a outra por funções multiplicativas ou exponenciais, a transformação logarítmica é uma das transformações de dados mais úteis (*ver* Nota de rodapé 13). Um exemplo ecológico clássico é a relação espécie-área: a relação entre o número de espécies e a área de ilhas ou de amostras (Preston, 1962; MacArthur & Wilson, 1967). Se medirmos o número de espécies em uma ilha e plotarmos contra a área da ilha, os dados com frequência seguem uma **função de potência** simples:

$$S = cA^z \tag{8.2}$$

Onde S é o número de espécies, A é a área da ilha, e c e z são constantes ajustadas aos dados. Por exemplo, o número de espécies de plantas encontradas em cada uma das ilhas de Galápagos (Preston, 1962) parece seguir uma relação assintótica (Tabela 8.2). Primeiro, observe que a área das ilhas varia mais de três ordens de grandeza, de menos de 1 a cerca de 7.500 km^2. Da mesma forma, o número de espécies varia em duas ordens de magnitude, de 7 a 325.

[15] Uma função contínua é uma função $f(X)$ tal que para quaisquer dois valores de uma variável aleatória X, x_i e x_j, que diferem por um número muito pequeno ($|x_i - x_j| < \delta$), então $|f(x_i) - f(x_j)| < \varepsilon$, outro número muito pequeno. Uma função monótona é uma função $f(X)$ tal que para quaisquer dois valores da variável aleatória X, x_i e x_j, se $x_i < x_j$ então $f(x_i) < f(x_j)$. Uma função monótona contínua possui essas duas propriedades.

TABELA 8.2 Riqueza de espécies em 17 ilhas de Galápagos

Ilha	Área (km^2)	Nº de espécies	$\log_{10}$(área)	$\log_{10}$(espécies)
Albemarle (Isabela)	5.824,9	325	3,765	2,512
Charles (Floreana)	165,8	319	2,219	2,504
Chatham (San Cristóbal)	505,1	306	2,703	2,486
James (Santiago)	525,8	224	2,721	2,350
Indefatigable (Santa Cruz)	1.007,5	193	3,003	2,286
Abingdon (Pinta)	51,8	119	1,714	2,076
Duncan (Pinzón)	18,4	103	1,265	2,013
Narborough (Fernandina)	634,6	80	2,802	1,903
Hood (Española)	46,6	79	1,669	1,898
Seymour	2,6	52	0,413	1,716
Barringon (Santa Fé)	19,4	48	1,288	1,681
Gardner 0.5	48	–0,286	1,681	
Bindloe (Marchena)	116,6	47	2,066	1,672
Jervis (Rábida)	4,8	42	0,685	1,623
Tower (Genovesa)	11,4	22	1,057	1,342
Wenman (Wolf)	4,7	14	0,669	1,146
Culpepper (Darwin)	2,3	7	0,368	0,845
Média	526,0	119,3	1,654	1,867
Desvio-Padrão	1.396,9	110,7	1,113	0,481
Erro-Padrão da Média	338,8	26,8	0,270	0,012

A área (originalmente apresentada em milhas quadradas) foi convertida para quilômetros quadrados e os nomes das ilhas usados por Preston (1962) foram mantidos, com os nomes modernos das ilhas apresentados entre parênteses. As duas últimas colunas mostram o logaritmo da área e da riqueza de espécies, respectivamente. Transformações matemáticas, como as logarítmicas, são usadas para atender melhor os pressupostos de análises paramétricas (normalidade, linearidade e variâncias constantes) e elas são usadas como uma ferramenta de diagnóstico para auxiliar na identificação de dados discrepantes e valores extremos. Como a variável resposta (riqueza de espécies) e a variável preditora (área da ilha) foram medidas em uma escala contínua, um modelo de regressão foi usado para a análise (*ver* Tabela 7.1).

Se plotarmos os dados brutos – o número de espécies S em função da área A (Figura 8.5) —, veremos que a maioria dos pontos está agrupada à esquerda da figura (pois a maioria das ilhas são pequenas). Como um primeiro passo para a análise, poderíamos tentar ajustar uma linha reta para essa relação:

$$S = \beta_0 + \beta_1 A \qquad (8.3)$$

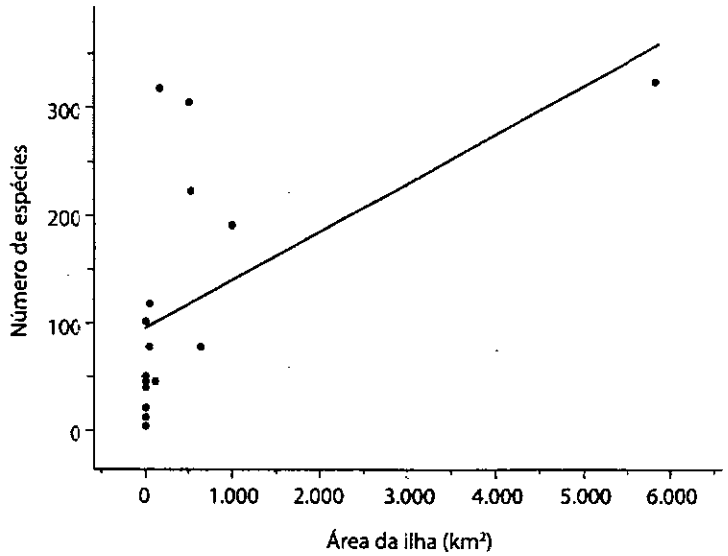


Figura 8.5 Gráfico da riqueza de espécies vegetais em função da área de ilhas de Galápagos utilizando os dados brutos (Colunas 2 e 3) da Tabela 8.2. A linha mostra o melhor ajuste da reta de regressão linear (*ver* Capítulo 9). Embora uma regressão linear possa ser ajustada para qualquer par de variáveis contínuas, o ajuste linear para esses dados não é bom: existem muitos dados discrepantes negativos nas ilhas pequenas, e a inclinação da reta é dominada pelos dados de Albermarle, a maior ilha do conjunto de dados. Em muitos casos, uma transformação matemática da variável *X*, da variável *Y*, ou de ambas irá melhorar o ajuste de uma regressão linear.

Neste caso, β_0 representa o intercepto da linha e β_1, sua inclinação (*ver* Capítulo 9). No entanto, a linha não se ajusta muito bem aos dados. Em particular, observe que a inclinação da linha de regressão parece ser dominada pelos dados de Albermarle, a maior ilha do conjunto de dados. No Capítulo 7, alertamos precisamente para esse problema, que pode aparecer quando dados discrepantes dominam o ajuste de uma linha de regressão (Figura 7.3). O ajuste da linha aos dados não captura bem a relação entre a riqueza de espécies e a área das ilhas.

Se a riqueza de espécies e a área das ilhas estão relacionadas exponencialmente (Equação 8.2), transformaremos essa equação tirando os logaritmos em ambos os lados:

$$\log(S) = \log(cA^z) \tag{8.4}$$

$$\log(S) = \log(c) + z\log(A) \tag{8.5}$$

Essa transformação tira vantagem de duas propriedades dos logaritmos. Primeiro, o logaritmo de um produto de dois números é igual à soma de seus logaritmos:

$$\log(ab) = \log(a) + \log(b) \tag{8.6}$$

Segundo, o logaritmo de um número elevado a uma potência é igual à potência multiplicada pelo logaritmo do número:

$$\log(a^b) = b\log(a) \tag{8.7}$$

Podemos reescrever a Equação 8.4, representando os valores transformados logaritmicamente com o símbolo *:

$$S^* = c^* + zA^* \tag{8.8}$$

Assim, tomamos uma equação exponencial (Equação 8.2) e a transformamos em uma equação linear (Equação 8.8). Quando plotamos o logaritmo dos dados, a relação entre riqueza de espécies e a área das ilhas agora é muito mais clara (Figura 8.6) e os coeficientes têm uma interpretação simples. O valor de z nas Equações 8.2 e 8.8, que equivale à inclinação da linha na Figura 8.6, é de 0,331. Isso significa que, aumentando A^* (o logaritmo da área da ilha) em uma unidade (ou seja, por um fator de 10 vezes, já que usamos o $\log_{10}$ para transformar nossas variáveis na Tabela 8.2, e $10^1 = 10$), aumentamos a riqueza de espécies em 0,331 unidades (ou seja, por um fator de aproximadamente 2, pois $10^{0,331}$ é igual a 2,14). Assim, podemos dizer que um aumento de 10 vezes na área

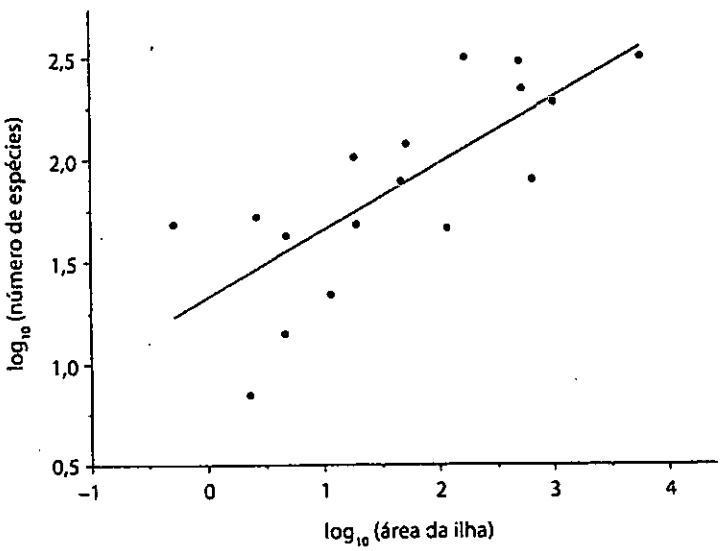


Figura 8.6 Gráfico do logaritmo da riqueza de espécies de plantas em função do logaritmo da área das ilhas de Galápagos (Colunas 4 e 5 da Tabela 8.2). A reta é a linha de regressão com melhor ajuste (*ver* Capítulo 9). Comparada com o gráfico da Figura 8.5, essa regressão se ajusta aos dados: a maior ilha do conjunto de dados não parece mais uma discrepância e a linearidade do ajuste é melhor.

das ilhas resulta em uma duplicação do número de espécies presentes nas ilhas de Galápagos.[16]

Outras transformações podem ser usadas para converter relações não lineares em lineares. Por exemplo, a transformação pela raiz cúbica ($\sqrt[3]{Y}$) é adequada para medidas de massa ou volume (Y^3), as quais relacionadas alometricamente com medidas lineares de tamanho ou comprimento corporal (Y; *ver* Nota de rodapé 13). Em estudos que examinam as relações entre duas medidas de massa ou volume (Y^3), como a comparação das massas cerebral e corporal, tanto a variável X quanto Y são logaritmicamente transformadas. A transformação logarítmica reduz a variação dos dados, que pode variar em várias ordens de magnitude (para uma exposição detalhada da relação entre a massa cerebral e a corporal, *ver* Allison & Cicchetti, 1976; Edwards, 2000).

Transformações de dados por demanda estatística

Todos os testes estatísticos requerem que os dados atendam a alguns pressupostos matemáticos. Por exemplo, os dados a serem analisados usando análise de variância (Capítulos 7 e 10) devem satisfazer dois pressupostos:

1. Os dados devem ser **homocedásticos** – ou seja, as variâncias dos resíduos de todos os grupos de tratamentos devem ser semelhantes entre si.
2. Os **resíduos**, ou desvios da média de cada grupo, devem ser variáveis aleatórias normais.

Da mesma forma, os dados a serem analisados usando regressão ou correlação (Capítulo 9) também devem ter resíduos com distribuição normal e que não sejam correlacionados com a variável independente.

[16] A história da relação espécie-área ilustra os perigos da **reificação**: a conversão de um conceito abstrato em uma coisa material. A função potência ($S = cA^z$) foi a base para vários modelos teóricos importantes da relação espécie-área (Preston, 1962; MacArthur & Wilson, 1967; Harte et al., 1999). Ela também tem sido usada como argumento para criar reservas naturais maiores, de forma que abriguem o maior número possível de espécies. Isso levou a um longo debate sobre se uma única reserva grande ou se várias reservas pequenas poderiam proteger a maioria das espécies (p. ex., Willis, 1984; Simberloff & Abele, 1984). No entanto, Lomolino e Weiser (2001) propuseram recentemente que a relação espécie-área tem uma assíntota, para a qual a função de potência não é apropriada. Mas essa proposta tem sido contestada com fundamentos teóricos e empíricos (Williamson et al., 2001). Em outras palavras, não há um consenso de que a função de potência sempre constitui a base para a relação espécie-área.

Essa função ficou popular porque ela parece fornecer um bom ajuste para muitas relações espécies-área. No entanto, uma análise estatística detalhada de 100 relações espécies-área publicadas (Connor & McCoy, 1979) constatou que a função potência foi o modelo com melhor ajuste em apenas metade dos conjuntos de dados. Embora a área das ilhas geralmente seja um preditor mais forte para o número de espécies, na maioria das vezes explica apenas metade da variação da riqueza de espécies (Boecklen & Gotelli, 1984). Como consequência, o valor da relação espécie-área para planejamentos de conservação é limitado, porque há muita incerteza associada às previsões de riqueza de espécies. Além disso, a relação espécie-área pode ser usada apenas para prever o número de espécies presentes, enquanto que a maioria das estratégias de manejo para conservação está interessada na identidade das residentes. A moral da história é que os dados sempre podem ser ajustados a uma função matemática, mas devemos usar ferramentas estatísticas para avaliar se o ajuste é razoável ou não. Mesmo se o ajuste dos dados for aceitável, o resultado, por si só, raras vezes é um teste forte para uma hipótese científica, pois em geral existem modelos matemáticos alternativos que bem podem se ajustar aos dados.

As transformações matemáticas dos dados podem ser usadas para atender esses pressupostos. Em geral, as transformações mais comuns visam a ambos os pressupostos de maneira simultânea. Em outras palavras, uma transformação que equaliza as variâncias (Pressuposto 1) frequentemente normaliza os resíduos (Pressuposto 2).

Cinco transformações são muito usadas com dados ecológicos e ambientais: a logarítmica, a da raiz quadrada, a angular (ou arcoseno), a recíproca e a de Box-Cox.

A TRANSFORMAÇÃO LOGARÍTMICA A **transformação logarítmica** (ou **transformação log**) substitui o valor de cada observação pelo seu logaritmo:[17]

$$Y^* = \log(Y) \tag{8.9}$$

A transformação logarítmica (na maioria das vezes usa o logaritmo natural ou logaritmo na base *e*)[18] em geral equaliza as variâncias de dados em que a média e a variância são positivamente correlacionadas. Uma correlação positiva

John Napier

[17] Os logaritmos foram inventados pelo escocês John Napier (1550-1617). Em seu tomo épico *Mirifici logarithmorum canonis descriptio* (1614), ele expõe a fundamentação para a utilização de logaritmos (a partir da tradução inglesa de 1616):

Vendo que não há nada (certamente aos apaixonados estudantes da matemática) que seja mais problemático na prática da matemática, nem que mais nos injuriam e dificultam nossos cálculos, do que as multiplicações, divisões, extrações quadradas e cúbicas de números grandes, que além do tedioso gasto de tempo estão em sua maioria sujeitos a muitos erros traiçoeiros, comecei, portanto, a considerar em minha mente qual artimanha poderia usar para remover esses entraves.

Como ilustrado nas Equações 8.4–8.8, o uso de logaritmos permite realizar multiplicação e divisão com simples adições e subtrações. Séries de números que são multiplicativas em uma escala linear se tornam aditivas em uma escala logarítmica (*ver* Nota de rodapé 2, Capítulo 3).

[18] Os logaritmos podem ser calculados em muitas "bases", e Napier não especificou uma base em particular em seu *Mirifici*. Em geral, para um valor a e uma base b, podemos escrever

$$\log_b a = X$$

que significa que $b^X = a$. Já usamos o logaritmo na "base 10" no exemplo da relação espécie-área na Tabela 8.2; para a primeira linha da Tabela 8.2, o $\log_{10}(7498) = 3{,}875$, pois $10^{3{,}875} = 7498$ (arredondando). Uma mudança em uma unidade $\log_{10}$ é uma potência de 10, ou uma ordem de grandeza. Outras bases logarítmicas usadas na literatura ecológica são a base 2 (as oitavas de Preston [1962] e usada para dados que aumentam em potências de 2), e a base e, a do logaritmo natural, ou aproximadamente 2,71828... e é um número transcendental, cujo valor foi comprovado por Leonhard Euler (1707–1783) pela igualdade

$$\lim_{n\to\infty}(1+\frac{1}{n})^n$$

O $\log_e$ foi inicialmente tratado como "logaritmo natural" pelo matemático Nicolaus Mercator (1620-1687), que não deve ser confundido com o cartógrafo Gerardus Mercator (1512-1594). Muitas vezes é escrito como ln em vez de $\log_e$.

significa que grupos com médias grandes também terão variâncias grandes (na ANOVA) ou que a magnitude dos resíduos está correlacionada com a da variável independente (na regressão). Os dados univariados que são positivamente assimétricos (assimetria à direita) muitas vezes contêm alguns dados discrepantes grandes. Com uma transformação de log, esses dados geralmente são atraídos para a parte principal da distribuição, tornando-se mais simétrica. Os conjuntos de dados que apresentam médias e variâncias positivamente correlacionadas também tendem a ter dados discrepantes com resíduos positivamente assimétricos. A transformação logarítmica com frequência resolve os dois problemas ao mesmo tempo.

Note que o logaritmo de 0 não é definido: independentemente da base b, não há nenhum número em que $a^b = 0$. Uma maneira de contornar esse problema é adicionar 1 a cada observação antes de tirar seu logaritmo (como o $\log(1) = 0$, independente da base b). No entanto, essa não é uma solução útil ou adequada, se o conjunto de dados (ou especialmente se um grupo de tratamento) contém muitos zeros.[19]

A TRANSFORMAÇÃO DA RAIZ QUADRADA A **transformação da raiz quadrada** substitui o valor de cada observação pela sua raiz quadrada:

$$Y^* = \sqrt{Y} \qquad \textbf{(8.10)}$$

Essa transformação com frequência é usada com dados de contagem – como o número de lagartas por paineirinha ou o número de *Rhexia* por município. No Capítulo 2, mostramos que esses dados muitas vezes seguem uma distribuição de Poisson e salientamos que a média e a variância de uma variável aleatória de Poisson são iguais (ao parâmetro λ da razão de Poisson). Assim, para uma variável aleatória de Poisson, a média e a variância variam de igual forma. Calcular a raiz quadrada de variáveis aleatórias de Poisson produz uma variância que é independente da média. Como a raiz quadrada de 0 é igual a 0, a transformação da raiz quadrada não muda dados que são iguais a 0. Assim, para completar a transformação, você deve somar um número pequeno aos valores antes de calcu-

[19] Adicionar uma constante antes de calcular logaritmos pode causar problemas para estimar a variabilidade da população (McArdle et al., 1990), mas isso não é uma questão séria para a maioria das estatísticas paramétricas. No entanto, existem algumas questões filosóficas sutis em como tratar os zeros em um conjunto de dados. Ao adicionar uma constante, você está dizendo que 0 representa um erro de medição. Presumivelmente, o verdadeiro valor é um número muito pequeno, mas, por chance, você tinha medido 0 naquela réplica. Contudo, se o valor for um zero real, a mais importante fonte de variação pode ser entre a ausência e a presença da quantidade medida. Neste caso, seria mais apropriado usar uma variável discreta com duas categorias para descrever o processo. Quanto mais zeros no conjunto de dados (e quanto mais forem concentrados em um tratamento), mais provável é que o 0 represente uma condição qualitativamente diferente do que uma medida positiva pequena. Em uma série temporal de uma população, um zero real representaria uma extinção da população em vez de apenas uma medida imprecisa.

lar a raiz quadrada. Sokal & Rohlf (1995) sugerem adicionar 1/2 (0,5) a cada valor, enquanto Anscombe (1948) sugere 3/8 (0,325).

A TRANSFORMAÇÃO DO ARCOSENO OU ARCOSENO DA RAIZ QUADRADA A **transformação do arcoseno, arcoseno da raiz quadrada** ou **transformação angular** substitui o valor de cada observação pelo arcoseno da raiz quadrada:

$$Y^* = \text{arcoseno}\sqrt{Y} \qquad \textbf{(8.11)}$$

Essa transformação é utilizada principalmente para proporções (e porcentagens), que são distribuídas como variáveis aleatórias binomiais. No Capítulo 3, observamos que a média de uma distribuição binomial = np e sua variância = $np(1 - p)$, onde p é a probabilidade de sucesso e n é o número de tentativas. Assim, a variância é uma função direta da média (variância = $(1 - p)$ vezes a média). A transformação do arcoseno (que é o inverso da função seno) remove essa dependência. Como a função seno gera apenas valores entre –1 e +1, o inverso da função seno pode ser aplicado apenas a dados cujos valores estão entre –1 e +1: $-1 \leq Y_i \leq +1$. Portanto, essa transformação é apropriada apenas para dados que são expressos como proporções (como p, a proporção ou a probabilidade de sucesso em uma tentativa binomial). Duas advertências sobre a transformação arcoseno. Primeiro, se seus dados são porcentagens (escala de 0 a 100), devem ser convertidos para proporções (escala de 0 a 1,0). Segundo, a função arcoseno gera dados transformados com unidades em radianos, não em graus.

A TRANSFORMAÇÃO INVERSA A **transformação inversa** substitui o valor de cada observação pelo seu valor inverso:

$$Y^* = 1/Y \qquad \textbf{(8.12)}$$

Normalmente é mais usada para dados que registram taxas, como o número de descendentes por fêmea. Em geral, dados de taxas parecem hiperbólicos quando plotados em uma função da variável no denominador. Por exemplo, se você plotar o número de descendentes por fêmea no eixo Y e o número de fêmeas na população no eixo X, a curva resultante pode parecer uma hipérbole, que decresce rapidamente no início, e depois conforme o X aumenta. A forma desses dados em geral é

$$aXY = 1$$

(onde X é o número de fêmeas e Y é o número de descendentes por fêmea), que pode ser re-escrita como uma hipérbole

$$1/Y = aX$$

Transformar o Y no seu valor inverso, $1/Y$, resulta em uma nova relação

$$Y^* = 1/(1/Y) = aX$$

que é mais sujeita a uma regressão linear.

A TRANSFORMAÇÃO BOX-COX Usamos o logaritmo, a raiz quadrada, o valor inverso e outras transformações para reduzir a variância e a assimetria nos dados e criar uma série de dados transformados que possuam aproximadamente uma distribuição normal. A última é a **transformação de Box-Cox,** ou transformação de potência generalizada. Essa transformação, na verdade, é uma família expressa pela equação

$$\begin{aligned} Y^* &= (Y^\lambda - 1)/\lambda \quad (\text{for } \lambda \neq 0) \\ Y^* &= \log_e(Y) \quad (\text{for } \lambda = 0) \end{aligned} \tag{8.13}$$

onde λ é o número que maximiza a **função log da verossimilhança:**

$$L = -\frac{\nu}{2}\log_e(s_T^2) + (\lambda - 1)\frac{\nu}{n}\sum_{i=1}^{n}\log_e Y \tag{8.14}$$

onde v são os graus de liberdade, n é o tamanho da amostra, e s_T^2 é a variância dos valores de Y transformados (Box & Cox, 1964). O valor de λ, que resulta quando a Equação 8.14 é maximizada, é usado na Equação 8.13 para fornecer o melhor ajuste dos dados transformados a uma distribuição normal. A Equação 8.14 deve ser resolvida iterativamente (tentando diferentes valores de λ até L ser maximizado), usando programas de computador.

Determinados valores de λ correspondem às transformações já descritas. Quando $\lambda = 1$, a Equação 8.13 resulta em uma transformação linear (operação de deslocamento; *ver* Figura 2.7); quando $\lambda = 1/2$, o resultado é a transformação da raiz quadrada; quando $\lambda = 0$, o resultado é a transformação logarítmica natural; e quando $\lambda = -1$, o resultado é a transformação inversa. Antes de partir para o problema de maximizar a Equação 8.14, você deve tentar transformar seus dados usando transformações aritméticas simples. Se seus dados têm assimetria à direita, tente usar as transformações mais familiares a partir da série $1/\sqrt{Y}, \sqrt{Y}, \ln(Y), 1/Y$. Se têm assimetria à esquerda, tente Y^2, Y^3, etc. (Sokal & Rohlf, 1995).

Apresentando resultados: transformados ou não?

Embora você possa transformar os dados para análise, você deve relatar os resultados nas unidades originais. Por exemplo, os dados de espécie-área da Tabela 8.2 poderiam ser analisados usando aqueles transformados em log; porém, ao descrever o tamanho da ilha ou a riqueza de espécies, você deve apresentá-los em suas

unidades originais, pela **transformação reversa**. Desta forma, o tamanho médio das ilhas será

$$\text{antilog}(\overline{\log A}) = \text{antilog}(1{,}654) = 45{,}110 \text{ km}^2$$

Note que isso é muito diferente da média aritmética do tamanho das ilhas antes da transformação, que era igual a 526 km^2. De forma similar, a riqueza média de espécies é

$$\text{antilog}(\overline{\log S}) = 10^{1{,}867} = 73{,}6 \text{ espécies}$$

que difere da média aritmética da riqueza de espécies antes da transformação, 119,3.

Você pode construir intervalos de confiança (*ver* Capítulo 3) de maneira similar, tirando o antilog dos limites de confiança construídos e usando os erros-padrão das médias dos dados transformados. Isso normalmente resulta em intervalos de confiança assimétricos. Para os dados da área das ilhas na Tabela 8.2, o erro-padrão da média do log(A) = 0,270. Para n = 17, o valor necessário a partir da distribuição t, $t_{0,025[16]} = 2{,}119$. Assim, o limite inferior do intervalo de confiança de 95% (usando a Equação 3.16) = 1,654 – 2,119 × 0,270 = 1,082. De forma similar, o limite superior do intervalo de confiança de 95% = 1,654 + 2,119 × 0,270 = 2,226. Em uma escala logarítmica, esses valores formam um intervalo simétrico em torno da média de 1,654, mas quando eles são retransformados, o intervalo não é mais simétrico. O antilog de 1,082 = $10^{1{,}082}$ = 12,08, enquanto o antilog de 2,226 = $10^{2{,}226}$ = 168.27, que não é simétrico em torno da média retransformada de 45,11.

O rastro de auditoria revisitado

O processo de transformação dos dados também deve ser adicionado ao seu rastro de auditoria. Assim como com os dados da GQ/CQ, os métodos utilizados para transformação de dados devem ser documentados nos metadados. Em vez de substituir seus dados na planilha original pelos valores transformados, você deve criar uma nova planilha. Essa planilha também deve ser armazenada em mídia permanente, receber um identificador único em seu catálogo de dados e ser adicionada ao rastro de auditoria.

RESUMO: O FLUXOGRAMA DE GERENCIAMENTO DE DADOS

A organização, a gestão e a curadoria de seu conjunto de dados são partes essenciais do processo científico e devem ser feitas antes de você começar a analisar seus dados. O processo completo está resumido em nosso fluxograma de gerenciamento de dados (Figura 8.7).

Conjuntos de dados bem organizados são uma contribuição duradoura para a comunidade científica e seu amplo compartilhamento aumenta a colaboração e

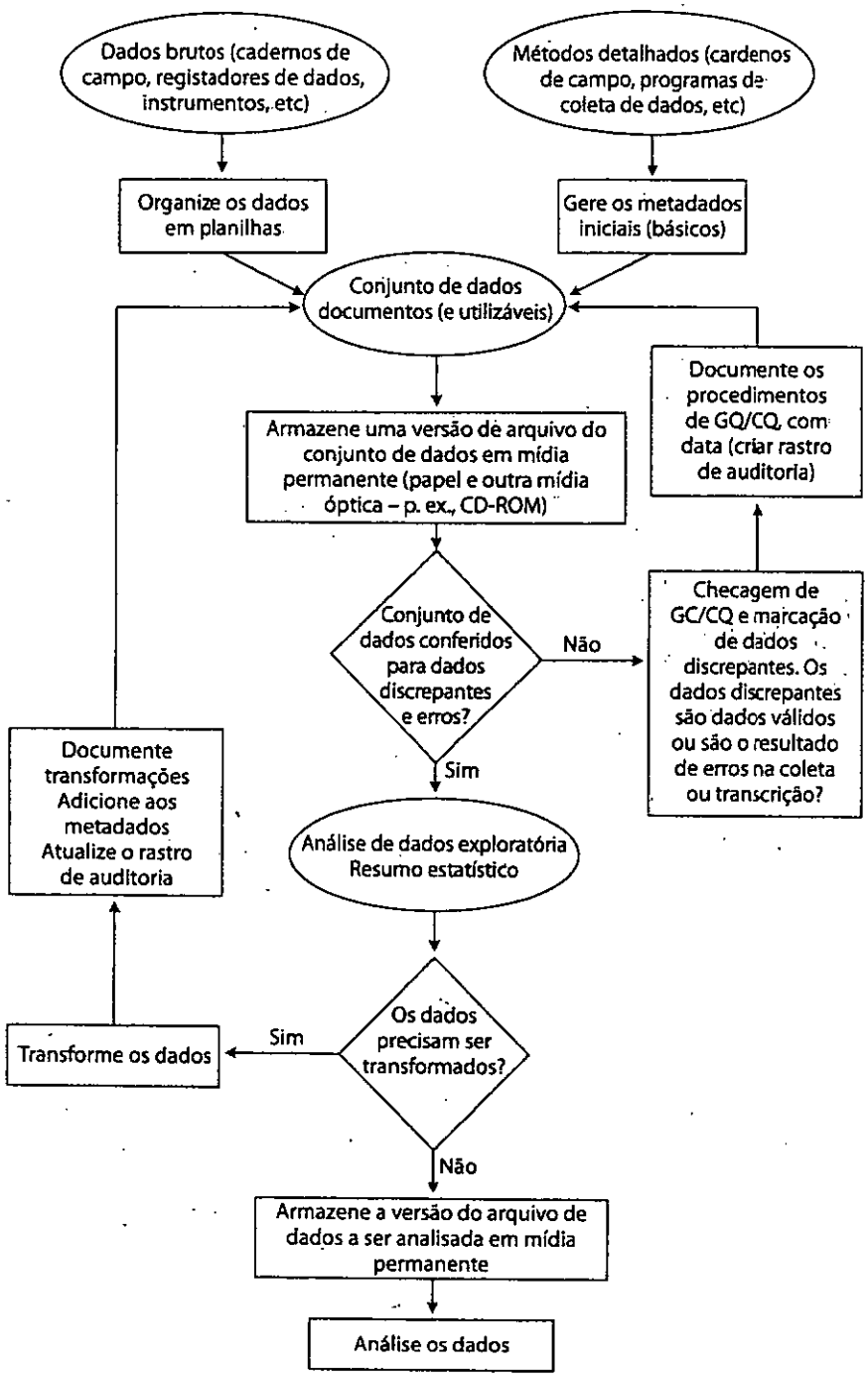


Figura 8.7 Fluxograma de gerenciamento de dados. Esse fluxograma descreve os passos para a gestão, a curadoria e o armazenamento de dados. Todos estes passos devem ser seguidos antes de a análise formal dos dados começar. A organização de metadados é especialmente importante para garantir que o conjunto de dados ainda será acessível e utilizável em um futuro distante.

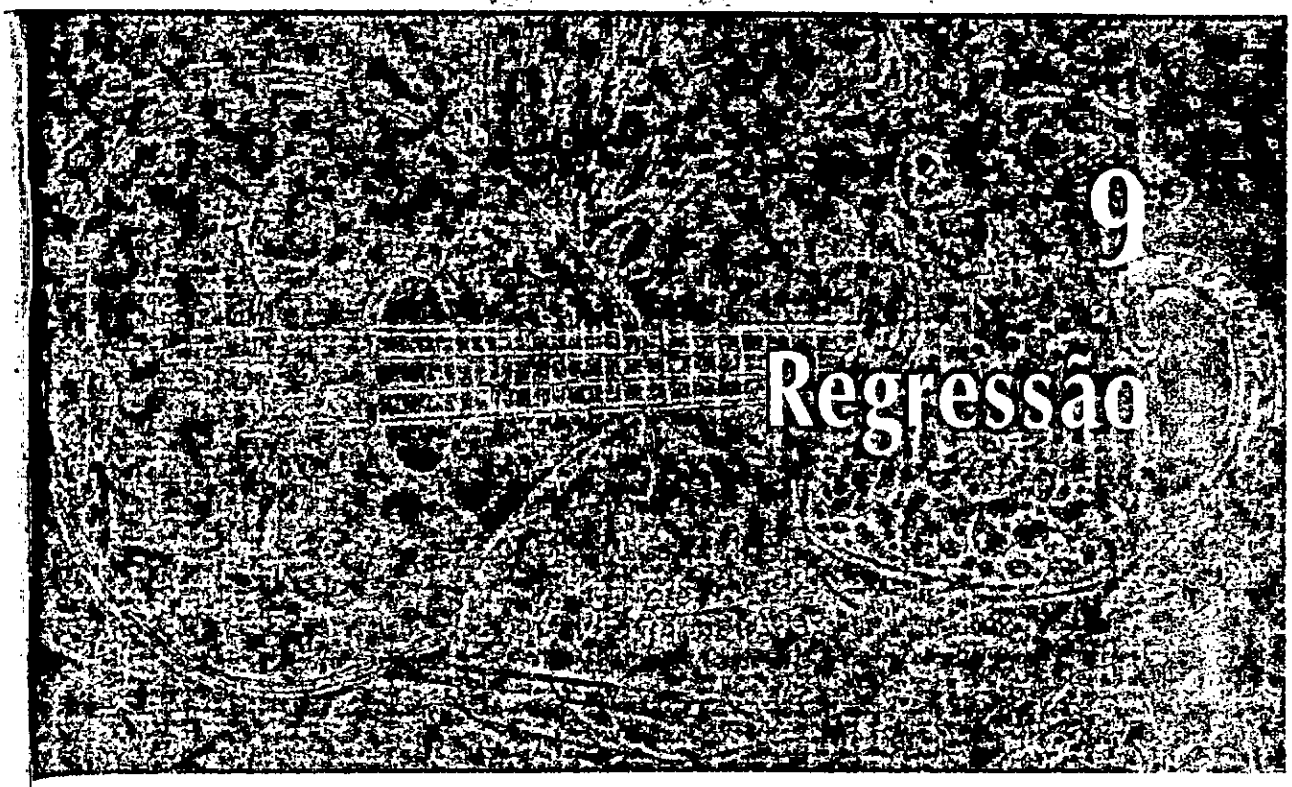
o ritmo do progresso científico. A disponibilidade livre e pública de dados é uma exigência de muitas agências de financiamento. Dados coletados com recursos públicos, distribuídos por agências e organizações nos Estados Unidos, podem ser solicitados a qualquer momento por qualquer pessoa, de acordo com as disposições da Lei de Liberdade de Informação. Os recursos financeiros adequados devem ser orçados para permitir o gerenciamento de dados.

Os dados devem ser organizados e informatizados rapidamente após a coleta, documentados com metadados necessários para reconstruir a coleta dos dados e arquivados em mídia permanente. Os conjuntos de dados devem ser conferidos com cautela em busca de erros de coleta e de transcrição, e em busca de dados discrepantes que não sejam dados válidos. Qualquer mudança ou alteração nos dados originais ("brutos") resultante de procedimentos de detecção de erros e dados discrepantes deve ser registrada detalhadamente nos metadados. Os arquivos alterados não devem substituir o de dados brutos. Ao contrário, devem ser armazenados como novos arquivos de dados (modificados). Um rastro de auditoria deve ser usado para procurar versões posteriores de conjuntos de dados e de seus metadados associados.

A análise de dados gráfica exploratória e o cálculo de estatísticas descritivas (p. ex., medidas de posição e de dispersão) devem preceder a análise estatística formal e os testes de hipóteses. Um exame cuidadoso dos gráficos e das tabelas preliminares pode indicar se os dados precisam ou não ser transformados antes de qualquer análise. As transformações dos dados também devem ser documentadas nos metadados. Uma vez que a versão final do conjunto de dados estiver pronta para análise, ela deverá ser arquivada em mídia permanente, com seus metadados e rastro de auditoria completos.

PARTE III

ANÁLISE DE DADOS



9 Regressão

A regressão é utilizada para analisar relações entre variáveis contínuas. Em sua forma mais básica, a regressão descreve a relação linear entre uma variável preditora, representada graficamente no eixo *X*, e uma variável resposta, representada no eixo *Y*. Neste capítulo, explicamos como o método de mínimos quadrados é usado para ajustar uma linha de regressão aos dados, e como testar a hipótese acerca dos parâmetros do modelo ajustado. Destacamos as premissas do modelo de regressão, descrevemos testes de diagnóstico para avaliar o ajuste dos dados ao modelo e explicamos como usar os modelos para fazer predições. Fornecemos, também, uma breve descrição de alguns tópicos mais avançados: regressão logística, regressão múltipla, regressão não linear, regressão robusta, regressão quantífica e análise de caminhos (*path analysis*). Terminamos com uma discussão sobre os problemas da seleção de modelos: como escolher um subconjunto de variáveis preditoras apropriado e como comparar o ajuste relativo de diferentes modelos ao mesmo conjunto de dados.

DEFININDO UMA LINHA RETA E SEUS DOIS PARÂMETROS

Começaremos com o desenvolvimento de um modelo linear, pois é o coração da análise de regressão. Como descrito no Capítulo 6, um modelo de regressão começa com a declaração de uma hipótese sobre causa e efeito: o valor da variável *X* causa, direta ou indiretamente, o valor da variável *Y*.[1] Algumas vezes, a direção

[1] Muitos textos de estatística enfatizam a distinção entre **correlação**, duas variáveis são meramente associadas uma com a outra, e **regressão**, uma relação direta de causa e efeito. Apesar de diferentes tipos de estatísticas terem sido desenvolvidas para ambos os casos, achamos que essa distinção é bastante arbitrária e, muitas vezes, apenas semântica. Afinal, os pesquisadores não procuram correlações entre variáveis a menos que acreditem ou suspeitem que exista alguma relação de causa e efeito subjacente. Neste capítulo, cobriremos os métodos estatísticos que são algumas vezes tratados de maneira individual nas análises de correlação e de regressão. Todas essas técnicas podem ser usadas para estimar e testar associações de duas ou mais variáveis contínuas.

da causa e efeito é direta – a hipótese de que a área das ilhas controla o número de espécies de plantas (Capítulo 8) –, e não o inverso. Em outros casos, a direção da causa e efeito não é tão óbvia – predadores controlam a abundância de suas presas ou a abundância de presas dita o número de predadores. (Capítulo 4)

Uma vez decidida a direção da causa e efeito, o próximo passo é descrever a relação como uma função matemática:

$$Y = f(X) \tag{9.1}$$

Em outras palavras, aplicamos a função f para cada valor da variável X (a entrada) para gerar o valor correspondente da variável Y (a saída ou o resultado). Existem muitas funções interessantes e complexas que podem descrever a relação entre duas variáveis, mas a mais simples é de que Y é uma função linear de X:

$$Y = \beta_0 + \beta_1 X \tag{9.2}$$

Em palavras, essa função diz "pegue o valor da variável X, multiplique por β_1 e some esse valor a β_0, o resultado é o valor da variável Y." Esta equação descreve o gráfico de uma linha reta. O modelo tem dois parâmetros, β_0 e β_1, que são chamados de **intercepto** e **inclinação** da reta (Figura 9.1). O intercepto (β_0) é o valor da função quando $X = 0$. O intercepto é medido nas mesmas unidades da variável Y. A inclinação é, portanto, uma taxa e é medida em unidade de $\Delta Y/\Delta X$ (leia-se: "mudança em Y dividida pela mudança em X"). Se a inclinação e o intercepto são conhecidos, a Equação 9.2 pode ser usada para predizer os valores de Y para qual-

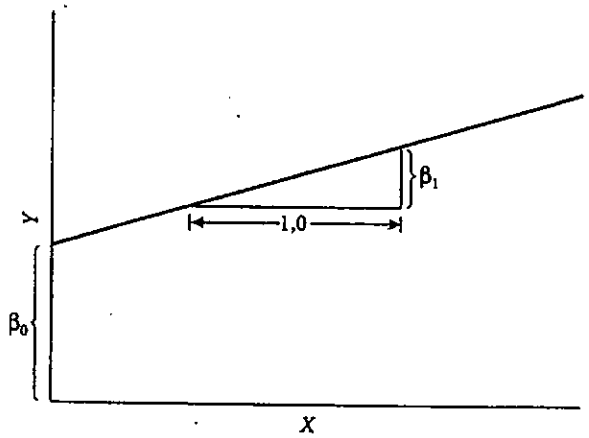


Figura 9.1 Relação linear entre as variáveis X e Y. A linha é descrita pela equação $Y = \beta_0 + \beta_1 X$, onde β_0 é o intercepto e β_1 é a inclinação da linha. O intercepto β_0 é o valor predito da equação quando $X = 0$. A inclinação da linha β_1 é o aumento na variável Y associado com o de uma unidade da variável X ($\Delta Y/\Delta X$). Se o valor de X é conhecido, o valor predito de Y pode ser calculado multiplicando X pela inclinação e somando o intercepto (β_0).

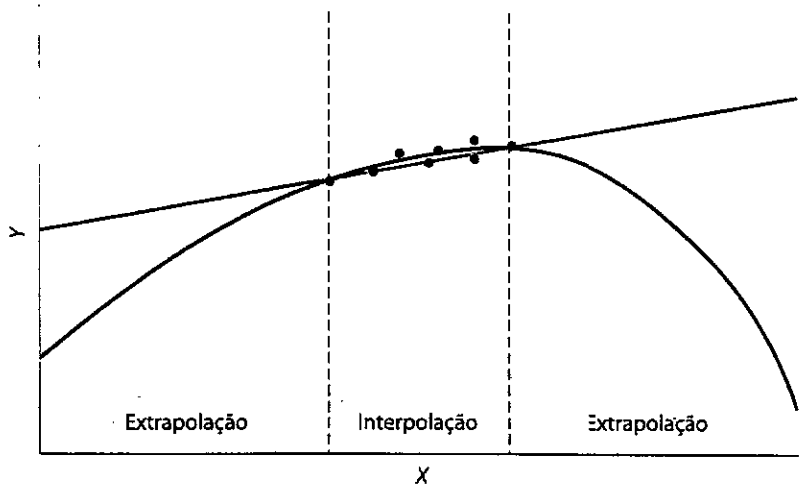


Figura 9.2 Modelos lineares podem aproximar funções não lineares sobre um domínio limitado da variável *X*. A interpolação dentro desses limites pode ser aceitável e acurada, embora o modelo linear (linha verde) não descreva a verdadeira relação funcional entre *Y* e *X* (curva preta). A extrapolação se tornará crescentemente menos acurada conforme as previsões se movem para além da amplitude dos dados coletados. Uma premissa muito importante da regressão linear é que a relação entre *X* e *Y* (ou transformações dessas variáveis) é linear.

quer valor de *X*. Reciprocamente, a Equação 9.2 pode ser usada para determinar o valor de *X*, podendo gerar um valor particular de *Y*.

A natureza não obedece uma equação linear, muitas relações ecológicas são inerentemente não lineares. Contudo, o modelo linear é o ponto de começo mais simples para ajustar funções aos dados. Além disso, mesmo as funções não lineares complexas (onduladas) podem ser *aproximadamente* lineares *sobre uma amplitude limitada da variável X* (Figura 9.2). Se, com cuidado, restringirmos nossas conclusões àquela amplitude da variável *X*, um modelo linear pode ser uma aproximação válida da função.

AJUSTANDO DADOS A UM MODELO LINEAR

Primeiro, vamos reformular a Equação 9.2 de forma que ela se adapte aos dados que coletamos. Os dados para uma análise de regressão consistem em uma série de observações pareadas. Cada observação inclui um valor de X (X_i) e um valor de Y correspondente (Y_i), ambos medidos para a mesma réplica. O subscrito i indica a réplica. Se houver um total de n réplicas em nossos dados, o subscrito i pode assumir qualquer valor inteiro de $i = 1$ até n. O modelo que iremos ajustar é

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (9.3)$$

Como na Equação 9.2, os dois parâmetros na equação linear, β_0 e β_1, são desconhecidos. Mas agora há uma terceira quantidade desconhecida, ε_i, que representa o termo do erro. Enquanto β_0 e β_1 são constantes simples, ε_i é uma variável aleatória normal. Esta distribuição tem um valor esperado (ou média) de 0, e variância igual a σ^2, que pode ser conhecida ou desconhecida. Se todos os dados caem perfeitamente em uma linha reta, então σ^2 é igual a 0, sendo uma tarefa fácil conectar esses pontos, e então, medir o intercepto (β_0) e a inclinação (β_1) diretamente a partir da linha.[2] Contudo, a maioria dos dados ecológicos exibe mais variação que isso – uma única variável raras vezes explicará a maior parte da variação dos dados, e os pontos dos dados cairão dentro de uma banda espalhada em vez de ao longo de uma linha acentuada. Quanto maior o valor de σ^2, mais ruído, ou erro, haverá sobre a linha de regressão.

No Capítulo 8, ilustramos uma linha de regressão para a relação entre a área das ilhas (a variável X) e o número de espécies de plantas (a variável Y) nas ilhas de Galápagos (Figura 8.6). Cada ponto na Figura 8.6 consistia em uma observação pareada da área da ilha e seu número de espécies de plantas correspondente (*ver* Tabela 8.2). Conforme explicamos no Capítulo 8, usamos uma transformação logarítmica tanto da variável X (área) quanto da Y (riqueza de espécies) para homogeneizar as variâncias e linearizar a curva. A seguir, neste capítulo, retornaremos a esses dados sem estarem transformados.

A Figura 8.6 mostra uma clara relação entre $\log_{10}$(área) e $\log_{10}$(número de espécies), mas os pontos não caem perfeitamente sobre uma linha. Onde a linha de regressão deve ser colocada? Intuitivamente, parece que deve passar através do centro da nuvem de dados, definida pelo ponto $(\overline{X}, \overline{Y})$. Para os dados das ilhas, o centro corresponde ao ponto (1,654, 1,867) (lembre-se de que são valores transformados).

Agora, podemos girar a linha através daquele ponto central até chegarmos à posição de "melhor ajuste". Mas como definimos o melhor ajuste para a linha? Primeiro, vamos definir o **resíduo quadrado** d_i^2 como a diferença quadrada entre o valor do Y_i observado e o do Y_i que é predito pela equação da regressão ($\hat{Y}_i$). Usamos o acento circunflexo para distinguir entre o valor observado (Y_i) e o valor predito a partir da equação de regressão ($\hat{Y}_i$). O resíduo quadrado d_i^2 é calculado como:

$$d_i^2 = (Y_i - \hat{Y}_i)^2 \qquad \textbf{(9.4)}$$

[2] É claro, se houver apenas duas observações, uma linha reta sempre irá se ajustar a elas perfeitamente. Mas com mais dados não há garantia de que uma linha reta seja um modelo significativo. Conforme veremos neste capítulo, uma linha reta pode ser ajustada a qualquer conjunto de dados e usada para fazer previsões, independente de quando o modelo ajustado é válido ou não. Contudo, com um grande conjunto de dados, temos ferramentas de diagnóstico para avaliar o seu ajuste e podemos estabelecer intervalos de confiança para as previsões derivadas do modelo.

O desvio é elevado ao quadrado porque estamos interessados na magnitude, e não no sinal, da diferença entre os valores observados e preditos (*ver* nota de rodapé 8, Capítulo 2). Para qualquer observação Y_i em particular, podemos passar a linha de regressão por aquele ponto, de forma que seu resíduo seja minimizado ($d_i = 0$). Mas a linha de regressão precisa se ajustar a todos os dados coletivamente. Portanto, definimos a soma de todos os resíduos, também chamada de **soma dos quadrados dos resíduos** e abreviada como *SQR**, como sendo:

$$SQR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (9.5)$$

A linha de regressão de "melhor ajuste" é a que minimiza a soma dos quadrados dos resíduos.[3] Ao minimizar a *SQR*, garantimos que a linha de regressão resulte na menor diferença média entre cada valor de Y_i do valor ($\hat{Y}_i$) que é predito pelo modelo de regressão (Figura 9.3).[4]

Francis Galton

[3] Francis Galton (1822-1911) – explorador inglês, antropólogo, primo de Charles Darwin, outrora estatístico, quintessencial D.W.E.M. (*Dead White European Male*) — é mais lembrado por seu interesse em eugenia e por sua afirmação de que a inteligência é herdada e pouco afetada por fatores ambientais. Seus escritos sobre inteligência, raça e hereditariedade o levaram a defender restrições reprodutivas entre pessoas, fundamentando políticas racistas na Austrália Colonial. Ele foi elevado a cavaleiro em 1909.

Em seu artigo de 1866, "Regression towards mediocrity in hereditary structure," Galton analisou a altura de filhos adultos e de seus pais com o modelo de regressão por mínimos quadrados. Apesar de os dados de Galton frequentemente serem usados para ilustrar a regressão linear e a regressão em direção à média, uma reanálise recente dos dados originais demonstrou que eles não são lineares (Wachsmuth et al., 2003).

Galton também inventou um dispositivo chamado de *quincunx* (tábua de Galton), uma caixa com face de vidro, com fileiras de pinos dentro e um afunilamento no topo. Um chumbo de tamanho e massa uniforme é jogado através da parte afunilada e defletido pelos pinos em direção a um número modesto de compartimentos. Jogue o suficiente e obterá uma curva normal (gaussiana). Galton usou esse dispositivo para obter dados empíricos, mostrando que a curva normal pode ser expressa como a mistura de outras curvas normais.

[4] Ao longo deste capítulo, apresentaremos fórmulas, como a Equação 9.5, para a soma dos quadrados e outras quantidades usadas em estatística. Contudo, não recomendamos que você use as fórmulas desta forma para fazer seus cálculos. Ocorre que essas fórmulas são muito suscetíveis a erros de arredondamento, que acumulam em uma soma grande (Press et al., 1986). A multiplicação de matrizes é uma forma muito mais confiável de obter soluções estatísticas. De fato, ela é a única maneira de solucionar problemas mais complexos, como a regressão múltipla. É importante que você estude essas equações para entender como a estatística funciona, e é uma boa ideia usar um *software* de planilha eletrônica para tentar alguns exemplos "a mão". Contudo, para análises e publicações, você deve deixar que um pacote dedicado à estatística faça os cálculos numéricos pesados.

* N. de T. *RSS* em inglês, para *Residual Sum of Squares*.

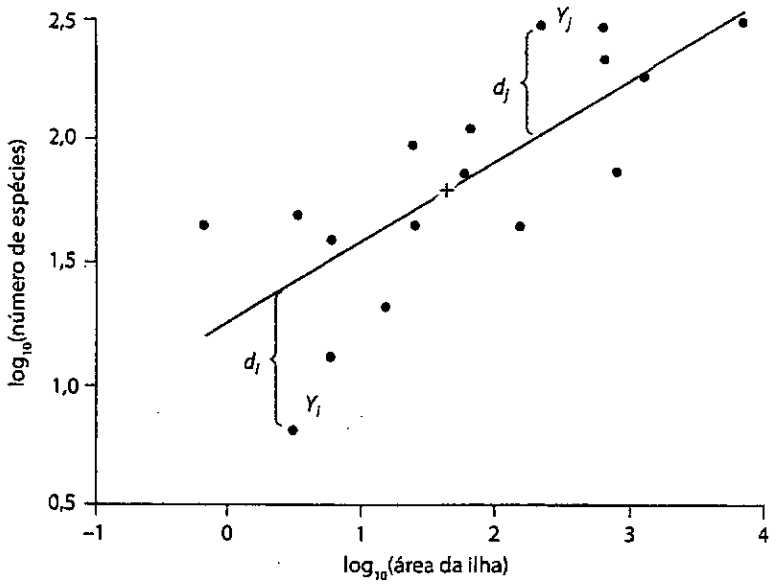


Figura 9.3 A soma dos quadrados dos resíduos é obtida somando os desvios quadrados (d_i) de cada observação da linha de regressão ajustada. A estimativa do parâmetro dos mínimos quadrados garante que essa linha de regressão minimize a soma dos quadrados dos resíduos. O + marca o ponto central dos dados ($\bar{X}$, $\bar{Y}$). Essa linha de regressão descreve, ainda, a relação entre o logaritmo da área das ilhas e o do número de espécies de plantas das ilhas de Galápagos, dados da Tabela 8.2. A equação da regressão é $\log_{10}(\text{espécies}) = 1{,}320 + \log_{10}(\text{área}) \times 0{,}331$; $r^2 = 0{,}584$.

Poderíamos ajustar a linha de regressão através de ($\bar{X}$, $\bar{Y}$) "no olho" e, então, mexer com ela até encontrarmos o valor da inclinação e do intercepto que nos dê o menor valor de *SQR*. Felizmente, existe uma forma mais simples de obter as estimativas de β_0 e de β_1 que minimizam a *SQR*. Porém, primeiro precisamos fazer um breve desvio para discutir variâncias e covariâncias.

VARIÂNCIAS E COVARIÂNCIAS

No Capítulo 3, apresentamos a soma dos quadrados (SQ_Y) de uma variável,

$$SQ_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \qquad (9.6)$$

que mede o desvio quadrado de cada observação a partir da média das observações. Dividindo essa soma por $(n - 1)$, temos a fórmula familiar para a variância amostral de uma variável:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{9.7}$$

Removendo o expoente da Equação 9.6 e expandindo o termo quadrado, podemos re-escrever como:

$$SQ_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y}) \tag{9.8}$$

Agora, consideremos a situação com duas variáveis X e Y. Em vez da soma dos quadrados de uma variável, podemos definir a **soma dos produtos cruzados** (SQ_{XY}) como:

$$SQ_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{9.9}$$

e a **covariância amostral** (s_{XY}) como:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{9.10}$$

Como vimos no Capítulo 3, a variância amostral é sempre um número não negativo. Porque o quadrado da diferença de cada observação a partir da média $(Y_i - \bar{Y})^2$ é sempre maior que zero e a soma de todas essas diferenças quadradas também deve ser. Mas o mesmo não é verdade para a covariância amostral. Suponha que valores relativamente grandes de X sejam consistentemente pareados com valores de Y relativamente pequenos. O primeiro termo na Equação 9.9 ou Equação 9.10 será positivo para os valores X_i que são maiores que $\bar{X}$. Mas o segundo termo (e também o produto) será negativo, pois os valores relativamente pequenos de Y_i são menores que $\bar{Y}$. De maneira similar, os valores relativamente pequenos de X_i (aqueles menores que $\bar{X}$) serão pareados com os relativamente grandes de Y_i (aqueles maiores que $\bar{Y}$). Se houver muitos pares de dados organizados desta forma, a covariância amostral será um número negativo.

Por outro lado, se os valores grandes de X estiverem sempre associados com valores grandes de Y, os termos de somatório das Equações 9.9 e 9.10 serão positivos, gerando um termo de covariância grande e positivo. Por fim, suponha que X e Y não são relacionados um ao outro, de forma que valores grandes ou pequenos de X possam ser, às vezes, associados com valores grandes ou pequenos de Y. Isso irá gerar uma coleção heterogênea de termos de covariância, com sinais negativos ou positivos. A soma desses termos pode ser próxima de zero.

Para os dados das plantas de Galápagos, $SQ_{XY} = 6{,}558$, e $s_{XY} = 0{,}410$. Todos os elementos do termo de covariância são positivos, exceto para as contribuições das ilhas Duncan (– 0,057) e Bindlow (– 0,080). Intuitivamente, pode parecer que essa medida de covariância deveria ser relacionada à inclinação da linha de regressão, porque ela descreve a relação (positiva ou negativa) entre a variação nas variáveis X e Y.[5]

ESTIMATIVA DE PARÂMETRO POR MÍNIMOS QUADRADOS

Tendo definido a covariância, podemos estimar os parâmetros da regressão que minimizam a soma dos quadrados dos resíduos:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{SQ_{XY}}{SQ_X} \tag{9.11}$$

onde a soma dos quadrados de X é:

$$SQ_X = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

Usamos o símbolo $\hat{\beta}_1$ para designar nossa estimativa da inclinação e para distingui-lo de β_1, o verdadeiro valor do parâmetro. Tenha em mente que β_1 tem um valor verdadeiro apenas na estrutura da estatística frequentista. Em uma análise bayesiana, os parâmetros por si só são vistos como uma amostra aleatória de uma distribuição (*ver* Capítulo 5). Depois, discutiremos a estimativa de parâmetros bayesianos em modelos de regressão.

A Equação 9.11 ilustra a relação entre a inclinação de um modelo de regressão e a covariância entre X e Y. Certamente, a inclinação é a covariância de X e Y,

[5] A covariância é um único valor que expressa a relação entre um único par de variáveis. Suponha que tenhamos um conjunto de n variáveis. Para cada par único (x_i, x_j) de variáveis podemos calcular cada um dos termos de covariância s_{ij}. Existem exatamente $n(n-1)/2$ pares únicos, que podem ser determinados usando o coeficiente binomial do Capítulo 2 (*ver*, também, a nota de rodapé 4, Capítulo 2). Esses termos de covariância podem, então, ser arranjados em uma **matriz de variância-covariância** quadrada $n \times n$, na qual todas as variáveis são representadas em cada linha e em cada coluna da matriz. A entrada da matriz na linha i, coluna j, é simplesmente s_{ij}. Os elementos da diagonal dessa matriz são as variâncias de cada variável $s_{ii} = s_i^2$. Os elementos da matriz acima e abaixo da diagonal são imagens-espelho uma da outra, pois, para cada par de variáveis X_i e X_j, a Equação 9.10 mostra que $s_{ij} = s_{ji}$. A matriz de variância-covariância é uma peça-chave do maquinário usado para obter as soluções de mínimos quadrados para muitos problemas estatísticos (*ver também*, o Capítulo 12 e o Apêndice).

A matriz de variância-covariância também faz uma aparição na ecologia de comunidades. Suponha que cada variável represente a abundância de uma espécie em uma comunidade. Intuitivamente, a covariância na abundância deveria refletir as formas de interações que ocorrem entre pares de espécies (negativa para predação e competição, positiva para mutualismo, 0 para interações fracas ou neutras). A **matriz da comunidade** (Levins, 1968) e outras medidas de coeficientes de interação par-a-par podem ser computadas a partir da matriz de variância-covariância das abundâncias e usadas para predizer a dinâmica e a estabilidade da assembleia inteira (Laska e Wootton, 1998).

na escala da variância de X. Como o denominador (n – 1) é idêntico para o cálculo de s_{XY} e de s_X^2, $\hat{\beta}_1$ também pode ser expresso como a razão da soma dos produtos cruzados (SQ_{XY}) pela soma dos quadrados de X (SQ_X). Para os dados das plantas de Galápagos, $s_{XY} = 0{,}410$, e $s_X^2 = 1{,}240$, então, $\hat{\beta}_1 = 0{,}410/1{,}240 = 0{,}331$. Lembre-se de que a inclinação sempre é expressa em unidades de $\Delta Y/\Delta X$, que, neste caso, é a mudança em $\log_{10}$(número de espécies) dividido pela mudança em $\log_{10}$(área da ilha).

Para encontrar a solução para o intercepto da equação ($\hat{\beta}_0$), tiramos vantagem do fato que a linha de regressão passa através do ponto ($\overline{X}$, $\overline{Y}$). Combinando isso à estimativa do $\hat{\beta}_1$, temos:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{9.12}$$

Para os dados de plantas das ilhas de Galápagos, $\hat{\beta}_0 = 1{,}867 - (0{,}331)(1{,}654) = 1{,}319$.

As unidades do intercepto são as mesmas da variável Y, que, neste caso, é $\log_{10}$(número de espécies). O intercepto nos dá a estimativa da variável resposta quando o valor da variável X é igual a zero. Para os nossos dados das ilhas, $X = 0$ corresponde a uma área de 1 km^2 (lembre-se de que o $\log_{10}(1{,}0) = 0{,}0$), com uma estimativa de $10^{1{,}319} = 20{,}844$ espécies.

Ainda há um último parâmetro para estimar. Na Equação 9.3, nosso modelo de regressão não incluiu apenas um intercepto ($\hat{\beta}_0$) e uma inclinação ($\hat{\beta}_1$), mas também o termo do erro (ε_i). O termo do erro tem distribuição normal com média 0 e variância σ^2. Como podemos estimar σ^2? Primeiro, se σ^2 for relativamente grande, os dados observados devem ser amplamente dispersos em torno da linha de regressão. Às vezes, a variável aleatória será positiva, puxando o dado Y_i acima da linha de regressão. Outros, negativa, puxando o Y_i abaixo da linha. Quanto menor for o σ^2, mais fortemente os dados irão se agrupar ao redor da linha de regressão ajustada. Por fim, se o $\sigma^2 = 0$, não deverá haver nenhuma dispersão, e os dados "cairão sobre a linha" perfeitamente.

Tal descrição parece muito similar à explicação para a soma dos quadrados dos resíduos (SQR), que mede o desvio quadrado de cada observação de seu valor ajustado (Equação 9.5). Na estatística descritiva (Capítulo 3), a variância amostral mede o desvio médio de quanto cada observação sai da média (Equação 3.9). De maneira similar, nossa estimativa da variância do erro da regressão é o desvio médio de quanto cada observação difere do valor ajustado:

$$\hat{\sigma}^2 = \frac{SQR}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right]^2}{n-2} \tag{9.13}$$

As fórmulas expandidas são mostradas para lembrá-lo do cálculo da SQR e do $\hat{Y}_i$. Como antes, relembre que $\hat{\beta}_0$ e $\hat{\beta}_1$ são os parâmetros ajustados da regressão, e $\hat{Y}_i$ é o valor predito a partir da equação da regressão. A raiz quadrada da Equação 9.13, $\hat{\sigma}$, em geral é chamada de **erro-padrão da regressão**. Note que o denominador da variância estimada é $(n-2)$, enquanto antes usávamos $(n-1)$ como denominador no cálculo da variância amostral (Equação 3.9). A razão para usar $(n-2)$ é que o denominador são os graus de liberdade, o número de informações independentes que temos para estimar aquela variância (ver Capítulo 3). Neste caso, já usamos dois graus de liberdade para estimar o intercepto e a inclinação da linha de regressão. Para os dados das plantas de Galápagos, $\hat{\sigma} = 0{,}320$.

COMPONENTES DE VARIÂNCIA E COEFICIENTE DE DETERMINAÇÃO

Uma técnica fundamental das análises paramétricas é a partição da soma dos quadrados em diferentes componentes, ou fontes. Apresentaremos a ideia aqui e retornaremos a ela no Capítulo 10. A partir dos dados brutos, consideraremos a soma dos quadrados da variável Y (SQ_Y) para representar a variação total da qual tentamos fazer a partição.

Um componente da variação total é puro ou erro aleatório. Essa variação não pode ser atribuída a nenhuma fonte em particular que não a amostragem aleatória de uma distribuição normal. Na Equação 9.3, essa fonte de variação é ε_i, e já vimos como estimar essa variação residual – calculando a soma dos quadrados dos resíduos (SQR). A variação remanescente em Y_i não é aleatória, mas sistemática. Alguns valores de Y_i são grandes *porque* estão associados a valores grandes de X_i. A fonte dessa variação é a relação da regressão $Y_i = \beta_0 + \beta_1 X_i$. Por subtração, o componente de variação remanescente que pode ser atribuído ao modelo de regressão (SQ_{reg}) é:

$$SQ_{reg} = SQ_Y - SQR \qquad (9.14)$$

Rearranjando a Equação 9.14, a variação total nos dados pode ser aditivamente dividida nos componentes da regressão (SQ_{reg}) e dos resíduos (SQR):

$$SQ_Y = SQ_{reg} + SQR \qquad (9.15)$$

Para os dados de Galápagos, $SQ_Y = 3{,}708$ e $SQR = 1{,}540$. Portanto, $SQ_{reg} = 3{,}708 - 1{,}540 = 2{,}168$.

Podemos imaginar dois extremos neste "corte da torta de variância". Suponha que todos os dados caiam com perfeição sobre a linha de regressão, de forma que qualquer valor de Y_i possa ser exatamente predito sabendo o valor do X_i. Assim,

$SQR = 0$ e $SQ_Y = SQ_{reg}$. Em outras palavras, toda a variação nos dados pode ser atribuída à regressão e não a um componente de erro aleatório.

Por outro lado, suponha que a variável X não tem nenhum efeito sobre a Y. Se não há nenhuma influência de X em Y, então $\beta_1 = 0$ e a linha não tem inclinação:

$$Y_i = \beta_0 + \varepsilon_i \tag{9.16}$$

Como ε_i é uma variável aleatória com média 0 e variância σ^2, tirando vantagem da operação de deslocamento (Capítulo 2), temos:

$$Y \sim N(\beta_0, \sigma) \tag{9.17}$$

Em palavras, a Equação 9.17 diz que Y é uma variável aleatória normal com média (ou expectativa) igual a β_0 e desvio-padrão σ. Se nada da variação em Y pode ser atribuído à regressão, a inclinação é igual a 0 e os Y_i são tirados de uma distribuição normal com média igual ao intercepto da regressão (β_0) e variância igual a σ^2. Neste caso, $SQ_Y = SQR$, portanto, $SQ_{reg} = 0{,}0$.

Entre os dois extremos de $SQ_{reg} = 0$ e $SQR = 0$ está a realidade da maioria dos conjuntos de dados, que reflete tanto a variação aleatória quanto a sistemática. Um índice natural que descreve a importância relativa da regressão *versus* a variação residual é o familiar r^2, ou **coeficiente de determinação**:

$$r^2 = \frac{SQ_{reg}}{SQ_Y} = \frac{SQ_{reg}}{SQ_{reg} + SQR} \tag{9.18}$$

O coeficiente de determinação informa a proporção de variação na variável Y pode ser atribuída à variação na variável X através de uma simples regressão linear. Essa proporção varia de 0,0 a 1,0. Quanto maior o valor, menor é a variância do erro e mais próximos os dados ficam da linha de regressão ajustada. Para os dados de Galápagos, $r^2 = 0{,}585$, cerca de meio caminho entre nenhuma correlação e o ajuste perfeito. Se você converter o valor do r^2 para uma escala de 0 a 100, o resultado obtido em geral é tratado como a porcentagem de variação em Y explicada pela regressão com o X. Lembre-se, contudo, que a relação causal entre as variáveis X e Y é uma hipótese proposta de maneira explícita pelo pesquisador (*ver* Equação 9.1). O coeficiente de determinação – não importa quão grande – não confirma por si só uma relação de causa e efeito entre as duas variáveis.

Uma estatística relacionada é o **coeficiente de correlação momento-produto**, r. Como você pode imaginar, o r é apenas a raiz quadrada do r^2. Contudo, o sinal do r (positivo ou negativo) é determinado pelo da inclinação da regressão:

negativo se o $\beta_1 < 0$, e positivo se o $\beta_1 > 0$. Equivalentemente, o r pode ser calculado como:

$$r = \frac{SQ_{XY}}{\sqrt{(SQ_X)(SQ_Y)}} = \frac{s_{XY}}{s_X s_Y} \tag{9.19}$$

com o sinal positivo ou negativo saindo do termo de soma dos produtos cruzados no numerador.[6]

TESTES DE HIPÓTESES COM REGRESSÃO

Até agora, aprendemos como ajustar uma linha reta a dados X e Y contínuos e como usar o critério de mínimos quadrados para estimar a inclinação, o intercepto e a variância da linha de regressão ajustada. O próximo passo é testar hipóteses acerca da linha de regressão ajustada. Relembre que o cálculo dos mínimos quadrados nos dá apenas estimativas ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$) dos valores reais dos parâmetros (β_0, β_1, σ^2). Como há incerteza nessas estimativas, precisamos testar se alguma difere significativamente de zero.

Em particular, o pressuposto subjacente da causa e efeito é incorporado no parâmetro de inclinação. Lembre-se de que, ao estipularmos um modelo de regressão, assumimos que X causa Y (*ver* Figura 6.2 para outras possibilidades). A magnitude do β_1 mede a força da resposta do Y às mudanças em X. Nossa hipótese nula é que β_1 não é diferente de zero. Se não pudermos rejeitar essa hipótese, não há nenhuma evidência compelindo uma relação funcional entre as variáveis X e Y. Estruturando a hipótese nula e a alternativa nos termos de nosso modelo, temos:

$$Y_i = \beta_0 + \varepsilon_i \quad \text{(Hipótese nula)} \tag{9.20}$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{(Hipótese alternativa)} \tag{9.21}$$

A anatomia de uma tabela ANOVA

A hipótese nula pode ser testada inicialmente organizando-se os dados em uma tabela de **análise de variância** (ANOVA). Apesar de a tabela de ANOVA ser natu-

[6] O Rearranjo da Equação 9.19 revela uma conexão íntima entre r e β_1, a inclinação da regressão:

$$\beta_1 = r\left(\frac{s_Y}{s_X}\right)$$

Assim, a inclinação da linha de regressão acaba por ser o coeficiente de correlação "re-escalado" em relação ao desvio-padrão do Y e do X.

ralmente associada à análise de variância (*ver* Capítulo 10), a partição da soma dos quadrados é associada à ANOVA, à regressão e a muitos outros modelos lineares generalizados (McCullagh e Nelder, 1989). A Tabela 9.1 ilustra uma tabela de ANOVA completa, com todos os seus componentes e suas equações. A Tabela 9.2 ilustra a mesma tabela de ANOVA com os cálculos para os dados das plantas de Galápagos. A forma abreviada da Tabela 9.2 é típica para uma publicação científica.

A tabela de ANOVA tem um número de colunas que sumariza a partição da soma dos quadrados. A primeira coluna, em geral, é intitulada "Fonte" e indica o componente ou a fonte de variação. No modelo de regressão existem apenas duas fontes: a regressão e o erro. Adicionamos uma terceira fonte, o total, para lembrá-lo de que a soma total dos quadrados é igual à dos quadrados da regressão e dos erros. Contudo, a linha referente à soma total dos quadrados geralmente é omitida das tabelas de ANOVA em publicações. Nosso modelo simples de regressão tem apenas duas fontes de variação, mas modelos mais complexos podem ter várias fontes de variação listadas.

A segunda coluna dá os graus de liberdade, em geral abreviados como gl*. Como havíamos discutido, os graus de liberdade dependem que quantas

TABELA 9.1 Tabela de ANOVA completa para uma regressão linear com um único fator

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)	Quadrado médio esperado	Razão-F	Valor de *P*
Régressão	1	$SQ_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$\frac{SQ_{reg}}{1}$	$\sigma^2 + \beta_1^2 \sum_{i=1}^{n} X^2$	$\frac{SQ_{reg}/1}{SQR/(n-2)}$	Cauda da distribuição F com 1 e $n-2$ graus de liberdade
Resíduos	$n-2$	$SQR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$\frac{SQR}{(n-2)}$	σ^2		
Total	$n-1$	$SQ_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$\frac{SQ_Y}{(n-1)}$	σ_Y^2		

A primeira coluna dá a fonte de variação nos dados. A segunda coluna dá os graus de liberdade (gl) associados a cada componente. Para uma regressão linear simples existe apenas 1 grau de liberdade associado à regressão e $(n-2)$ associados com o resíduo. O grau de liberdade total é $(n-1)$, pois 1 grau de liberdade sempre é usado para estimar a grande média dos dados. O modelo de regressão com um fator divide a variação total em um componente explicado pela regressão e em resíduo remanescente ("não explicado"). A soma dos quadrados é calculada usando os valores de Y observados (Y_i), a média dos valores de Y, ($\bar{Y}$), e os valores de Y preditos pelo modelo de regressão linear ($\hat{Y}_i$). Os quadrados médios esperados são usados para construir uma razão-F para testar a hipótese nula de que a variação associada com o termo da inclinação (β_1) é igual a 0,0. O valor do P para esse teste é retirado de uma tabela padrão de valores de F com 1 grau de liberdade para o numerador e $(n-2)$ graus de liberdade para o denominador. Os elementos básicos (fonte, gl, soma dos quadrados, quadrado médio, quadrado médio esperado, razão-F e o valor de P) são comuns a todas as tabelas de ANOVA em regressões e em análises de variância (Capítulo 10).

* N. de T. Em inglês df, para abreviar *degrees of freedom*.

TABELA 9.2 Forma de publicação da tabela de ANOVA para uma análise de regressão dos dados de espécie-área das plantas de Galápagos

Fonte	gl	*SQ*	*QM*	Razão-F	*P*
Regressão	1	2,168	2,168	21,048	0,000329
Resíduos	15	1,540	0,103		

Os dados brutos estão na Tabela 8.3, as fórmulas para esses cálculos estão na Tabela 9.1. A soma dos quadrados totais e os graus de liberdade raras vezes são incluídos em uma tabela de ANOVA publicada. Em muitas publicações, tais resultados poderiam ser reportados mais compactamente como "A regressão linear do log da riqueza de espécies pelo log da área foi altamente significante (modelo $F_{1,15} = 21{,}048$, $p < 0{,}001$). Assim, a equação de melhor ajuste foi $\log_{10}$ (Riqueza de espécies) = 1,867 + 0,331 × log (área das ilhas), $r^2 = 0{,}585$." As estatísticas de regressões publicadas devem incluir o coeficiente de determinação (r^2) e as estimativas de mínimos quadrados da inclinação (0,331) e do intercepto (1,867).

informações independentes estão disponíveis para estimar a soma dos quadrados em questão. Se o tamanho amostral é n, há um grau de liberdade associado ao modelo de regressão (especificamente a inclinação) e $(n-2)$ graus de liberdade associados ao erro. O grau de liberdade total é $(1 + n - 2) = (n - 1)$. O total é apenas $(n - 1)$ porque 1 grau de liberdade foi usado na estimativa da grande média, $\bar{Y}$.

A terceira coluna dá a soma dos quadrados (*SQ*) associada à fonte de variação em particular. Na Tabela 9.1 expandida, mostramos as fórmulas usadas, mas em tabelas publicadas (p. ex., Tabela 9.2) somente os valores numéricos para a soma dos quadrados seriam fornecidos.

A quarta coluna dá o quadrado médio (*QM*), que é a soma dos quadrados dividida por seus graus de liberdade correspondentes. Essa divisão é análoga ao cálculo de uma simples variância dividindo SQ_Y por $(n - 1)$.

A quinta coluna refere-se o quadrado médio esperado. Essa coluna não é apresentada em tabelas de ANOVA publicadas, mas é muito preciosa por mostrar exatamente o que está sendo estimado por cada um dos diferentes quadrados médios. Tais expectativas são usadas para formular testes de hipóteses na ANOVA.

A sexta coluna é a razão-F calculada (é a razão entre dois valores de quadrados médios diferentes).

A última coluna dá o valor de probabilidade em cauda correspondente a uma razão-F em particular. Especificamente, esta é a probabilidade de obter a razão-F observada (ou uma maior) se a hipótese nula for verdadeira. Para o modelo de regressão linear simples, a hipótese nula é que $\beta_1 = 0$, implicando em nenhuma relação funcional entre as variáveis X e Y. O valor da probabilidade depende do tamanho da razão-F e do número de graus de liberdade associados ao quadrado médio do numerador e ao do denominador. Os valores de probabilidade podem ser consultados em tabelas estatísticas, mas eles geralmente são parte dos resultados padrão retornados pelos pacotes estatísticos.

A fim de compreender como a razão-F é construída, precisamos examinar os valores dos quadrados médios esperados. O da regressão é a soma da variância do erro da regressão *e* um termo que mede o efeito da inclinação da regressão:

$$E(QM_{reg}) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} X^2 \qquad (9.22)$$

Em contraste, o valor esperado para o quadrado médio do resíduo é simplesmente a variância do erro da regressão:

$$E(QM_{resid}) = \sigma^2 \qquad (9.23)$$

Agora podemos entender a lógica por trás da construção da razão-F. A razão-F para o teste de regressão usa o quadrado médio da regressão no numerador e o quadrado médio dos resíduos, no denominador. Se a inclinação real da regressão é zero, o segundo termo na Equação 9.22 (o numerador da razão-F) também será zero. Como consequência, as Equações 9.22 (o numerador da razão-F) e 9.23 (o denominador da razão-F) serão iguais. Em outras palavras, se a inclinação da regressão (β_1) for zero, o valor esperado da razão-F será 1,0. Para uma dada quantidade de variância do erro: quanto mais acentuada a inclinação, maior a razão-F. Isto também tem sentido intuitivo, porque quanto menor a variância do erro, mais fortemente os dados estarão agrupados em torno da linha de regressão ajustada.

A interpretação do valor de *P* segue o raciocínio desenvolvido no Capítulo 4. Quanto maior a razão-F (para um dado tamanho amostral e modelo), menor o valor de *P*. Quanto menor o valor de *P*, mais improvável será que a razão-F observada seja encontrada caso a hipótese nula seja verdadeira. Com valores de *P* menores que o ponto de corte padrão de 0,05, rejeitamos a hipótese nula e concluímos que o modelo de regressão explica mais variação do esperado apenas ao acaso. Para os dados de Galápagos, a razão-F = 21,048. O numerador para essa razão-F é 21 vezes maior que o denominador, então a variância explicada pela regressão é muito maior que a residual. O valor de *P* correspondente = 0,0003.

Outros testes e intervalos de confiança

Como você pode suspeitar, todos os testes de hipóteses e intervalos de confiança para os modelos de regressão dependem da variância da regressão ($\hat{\sigma}^2$). A partir disso, podemos calcular outras variâncias e testes de significância (Weisberg, 1980). Por exemplo, a variância do intercepto estimado é:

$$\hat{\sigma}^2_{\beta_0} = \hat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{X}^2}{SQ_X}\right) \tag{9.24}$$

Uma razão-F também pode ser construída a partir dessa variância para testar a hipótese nula de que $\beta_0 = 0{,}0$.

Note que o intercepto da linha de regressão é subitamente diferente do que é calculado quando o modelo tem inclinação zero (Equação 9.16):

$$Y_i = \beta_0 + \varepsilon_i$$

Se o modelo tem uma inclinação 0, o valor esperado do intercepto é simplesmente a média dos valores de Y_i:

$$E(\beta_0) = \overline{Y}$$

Contudo, para o modelo de regressão com um termo de inclinação, o intercepto é o valor esperado quando $X_i = 0$, ou seja,

$$E(\beta_0) = \hat{Y}_i \mid (X_i = 0)$$

Um intervalo de confiança de 95% do tipo simples pode ser calculado para o intercepto como:

$$\hat{\beta}_0 - t_{(\alpha,n-2)}\hat{\sigma}_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + t_{(\alpha,n-2)}\hat{\sigma}_{\hat{\beta}_0} \tag{9.25}$$

onde α é o nível (duas caudas) de probabilidade ($\alpha = 0{,}025$ para um intervalo de confiança de 95%), n é o tamanho amostral, t é o valor de uma tabela de distribuição-t para o α e n especificados, e $\hat{\sigma}_{\hat{\beta}_0}$ é calculado como a raiz quadrada da Equação 9.24.

Similarmente, a variância da estimativa da inclinação é:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{SQ_X} \tag{9.26}$$

e o intervalo de confiança correspondente para a inclinação é:

$$\hat{\beta}_1 - t_{(\alpha,n-2)}\hat{\sigma}_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{(\alpha,n-2)}\hat{\sigma}_{\hat{\beta}_1} \tag{9.27}$$

Para os dados de Galápagos, o intervalo de confiança de 95% para o intercepto (β_0) é 1,017 a 1,623, e para a inclinação (β_1) é 0,177 a 0,484 (lembre-se de que esses são valores em $\log_{10}$). Como nenhum desses intervalos de confiança engloba o 0,0, o teste-F correspondente poderia nos fazer recusar a hipótese

nula de que β_0 e β_1 são iguais a 0. Se β_1 for igual a 0, a linha de regressão é horizontal e a variável dependente [$\log_{10}$(número de espécies)] não aumenta sistematicamente com mudanças na variável independente [$\log_{10}$(área das ilhas)]. Se β_0 for igual a 0, a variável dependente assume o valor de 0 quando a independente for 0.

Apesar de a hipótese nula ter sido rejeitada neste caso, os dados observados não caem perfeitamente sobre uma linha reta. Portanto, existe incerteza associada a qualquer valor à variável X em particular. Por exemplo, se você amostrar repetidamente diferentes ilhas com áreas idênticas (X), poderá haver variação entre elas no número de espécies (transformado em log) registrado.[7] A variância dos valores ajustados de Y é:

$$\hat{\sigma}^2_{(\hat{Y}|X)} = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SQ_X}\right) \tag{9.28}$$

e o intervalo de confiança de 95% é:

$$\hat{Y} - t_{(\alpha,n-2)}\hat{\sigma}_{(\hat{Y}|X)} \leq \hat{Y} \leq \hat{Y} + t_{(\alpha,n-2)}\hat{\sigma}_{(\hat{Y}|X)} \tag{9.29}$$

O intervalo de confiança não forma uma banda paralela que engloba a linha de regressão (Figura 9.4). Pelo contrário, o intervalo se abre quando mais nos distanciamos do $\bar{X}$, por causa do termo $(X_i - \bar{X})^2$ no numerador da Equação 9.28. Esse intervalo de confiança que se amplia tem sentido intuitivo. Quanto mais próximos estamos do centro da nuvem de pontos, mais confiança temos na estimativa do Y a partir de amostras com valores repetidos de X. De fato, se escolhermos $\bar{X}$ como o valor ajustado, o desvio-padrão do valor ajustado é equivalente a um erro-padrão do valor ajustado médio (*ver* Capítulo 3):

[7] Infelizmente, existe um desencontro entre a estrutura estatística de amostragem aleatória e a natureza de nossos dados. Afinal de contas, existe apenas um arquipélago de Galápagos, que evoluiu uma flora e fauna única, e não há múltiplas ilhas de réplica idênticas em área. Parece duvidoso tratar essas ilhas como amostras de um espaço amostral maior, que por si só não é claramente definido – poderiam ser ilhas vulcânicas? Ilhas tropicais do Pacífico? Ilhas oceânicas isoladas? Podemos pensar nos dados como uma amostra de Galápagos, exceto que a amostra dificilmente é aleatória, pois eles consistem de todas as grandes ilhas principais do arquipélago. Existem poucas ilhas adicionais das qual poderíamos coletar dados, mas elas são consideravelmente menores e com pouquíssimas espécies de plantas e animais. Algumas ilhas são tão pequenas que são essencialmente vazias, e esses dados de 0 (dos quais não podemos simplesmente tirar logaritmos) têm efeitos importantes na forma da relação (Williams, 1996). A linha de regressão ajustada para todas as ilhas pode não necessariamente ser a mesma para conjuntos de ilhas grandes *versus* pequenas. Tais problemas não são exclusivos de dados de relações espécie-área. Em qualquer estudo amostral precisamos nos deparar com o fato de que o espaço amostral não é claramente definido e que as réplicas coletadas podem não ser nem aleatórias nem independentes umas das outras, apesar de usarmos estatísticas que dependem desses pressupostos.

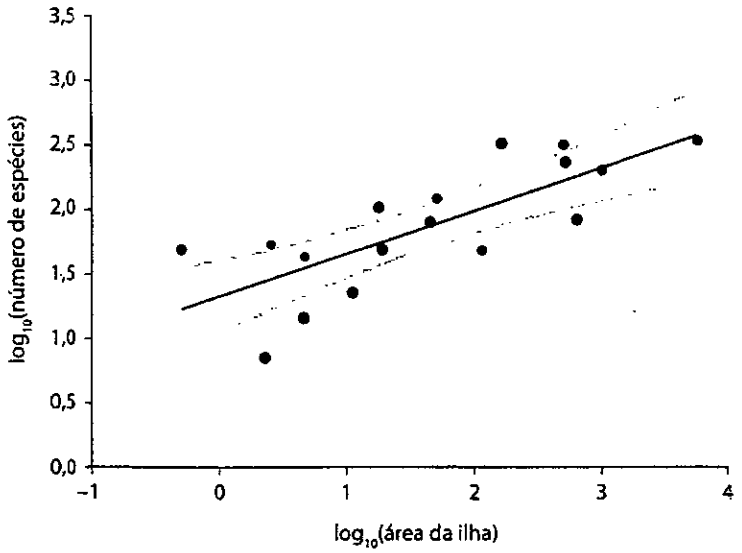


Figura 9.4 Linha de regressão (verde escuro), intervalo de confiança de 95% (linhas verdes) e intervalo de predição de 95% (linhas verde claro) para uma regressão log-log do número de espécies de plantas de Galápagos contra a área das ilhas (Tabela 8.2). O intervalo de confiança descreve a incerteza nos dados coletados, enquanto o intervalo de predição é usado para avaliar novos dados ainda não coletados ou incluídos nas análises. Note que o intervalo de confiança se abre conforme nos distanciamos da área média das ilhas com dados coletados. O intervalo de confiança se abrindo reflete maior incerteza acerca das predições para variáveis X, que são muito diferentes daquelas já coletadas. Note, também, que esses intervalos são descritos em uma escala logarítmica. A transformação reversa (e^Y) é necessária para converter as unidades em números de espécies. O intervalo de confiança com transformação reversa é muito largo e assimétrico, em torno dos valores previstos.

$$\hat{\sigma}_{(\hat{Y}|X)} = \hat{\sigma}/\sqrt{n}$$

A variância é minimizada aqui, pois existem dados observados tanto acima quanto abaixo do valor ajustado. Contudo, conforme nos distanciamos de $\bar{X}$, existem menos dados na vizinhança, portanto, a predição se torna menos confiável.

Uma distinção útil entre **interpolação** e **extrapolação** pode ser feita agora. A interpolação é a estimativa de novos valores que estão dentro da amplitude dos dados que coletamos; a extrapolação significa estimar novos valores além da amplitude dos dados (Figura 9.2). A Equação 9.29 garante que o intervalo de confiança para valores ajustados sempre será menor para a interpolação que para a extrapolação.

Por último, suponha que descubramos uma nova ilha no arquipélago, ou amostremos uma ilha que não foi indicada no censo original. Essa nova observação é designada como:

$$(\tilde{X}, \tilde{Y})$$

Usaremos a linha de regressão ajustada para construir um **intervalo de predição** para o novo valor, que é muito diferente de construir um intervalo de confiança para o valor ajustado. O intervalo de confiança engloba a incerteza acerca de uma estimativa com base nos dados disponíveis, enquanto o intervalo de predição engloba a incerteza acerca de uma estimativa fundamentada em novos dados. A variância da predição para um único valor ajustado é:

$$\sigma^2_{(\tilde{Y}|\tilde{X})} = \hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(\tilde{X}-\overline{X})^2}{SQ_X}\right) \qquad (9.30)$$

E o intervalo de predição correspondente é:

$$\tilde{Y} - t_{(\alpha,n-2)}\hat{\sigma}_{(\tilde{Y}|\tilde{X})} \leq \tilde{Y} \leq \tilde{Y} + t_{(\alpha,n-2)}\hat{\sigma}_{(\tilde{Y}\tilde{X})} \qquad (9.31)$$

Esta variância (Equação 9.30) é maior que a existente para o valor ajustado (Equação 9.28). A Equação 9.30 inclui tanto a variabilidade do erro associado à nova observação quanto a da incerteza na estimativa dos parâmetros da regressão dos dados iniciais.[8]

Para fechar essa sessão, notamos que é muito fácil (e sedutor) fazer predições pontuais a partir de uma linha de regressão. Contudo, para a maioria dos tipos de dados ecológicos, a incerteza associada com a predição em geral é muito grande para ser útil. Por exemplo, se propusermos uma reserva natural de 10 km^2 com base somente na informação contida na relação espécie-área de Galápagos, a predição pontual a partir da linha de regressão é de 45 espécies. Contudo, o intervalo de previsão de 95% para essa previsão é de 9 a 229 espécies. Se aumentarmos a área 10 vezes, para 100 km^2, a predição pontual seria de 96 espécies, podendo variar de 19 a 485 espécies (amplitudes com base em um intervalo de predição com transformação reversa a partir da Equação 9.31).[9] Esse intervalo de predição amplo não é típico apenas para dados de espécie-área, mas também para a maioria dos dados ecológicos.

PRESSUPOSTOS DA REGRESSÃO

O modelo de regressão linear que desenvolvemos até aqui recai sobre quatro pressupostos:

[8] Um refinamento final é que a Equação 9.31 é apropriada somente para uma predição única. Se múltiplas predições forem feitas, o valor do α precisa ser ajustado para criar um intervalo de predições simultâneas ligeiramente mais amplo. Também existem fórmulas para intervalos de predição inversa, em que uma amplitude de valores de X é obtida para um único valor da variável Y (*ver* Weisberg, 1980).

[9] Para ser exigente, a transformação reversa pontual estima a mediana, e não a média, da distribuição. Contudo, intervalos de confiança com transformação reversa não são tendenciosos, pois os quantis (pontos de porcentagem) se traduzem nas escalas transformadas.

1. **O modelo linear descreve corretamente a relação funcional entre *X* e *Y*.** Este é o pressuposto fundamental. Mesmo se a relação geral for não linear, um modelo linear pode continuar apropriado sobre uma amplitude limitada da variável *X* (Figura 9.2). Se o pressuposto de linearidade for violado, a estimativa do σ^2 será inflada, porque ele irá incluir tanto um erro aleatório quanto um componente de erro fixo (o último representa a diferença entre a função real e a linear que foi ajustada aos dados). E, se a relação real for não linear, as predições derivadas a partir do modelo serão enganosas, particularmente quando extrapoladas para além da amplitude dos dados.
2. **A variável *X* é medida sem erros.** Este pressuposto nos permite isolar o componente do erro inteiramente como variação aleatória associada com a variável resposta (*Y*). Se houver erros na variável *X*, as estimativas da inclinação e do intercepto serão tendenciosas. Assumindo que não há nenhum erro na variável *X*, podemos usar os estimadores de mínimos quadrados, que minimizam a distância vertical entre cada observação e seu valor predito (os d_i na Figura 9.3). Com erros tanto na variável *X* quanto na *Y*, uma estratégia seria minimizar a distância perpendicular entre cada observação e a linha de regressão. Chamada de regressão Modelo II, ela é comumente utilizada em componentes principais e outras técnicas multivariadas (*ver* Capítulo 12). Contudo, como a solução por mínimos quadrados tem se provado tão eficiente e é tão usada, esse pressuposto com frequência é ignorado.[10]
3. **Para qualquer dado valor de *X*, os valores de *Y* amostrados são independentes e com erros com distribuição normal.** O pressuposto de normalidade nos permite usar a teoria paramétrica para construir intervalos de confiança e testes de hipóteses com base na razão-F. Independência, obviamente, é o pressuposto crítico para todos os dados amostrais (*ver* Capítulo 6), mesmo sabendo que ela com frequência é violada a um grau desconhecido em estudos observacionais. Se você suspeitar que o valor Y_i influencia a próxima observação que coletar (Y_{i+1}), uma análise de séries temporais pode remover os componentes correlacionados da variação do erro. A análise de séries temporais foi discutida com brevidade no Capítulo 6.
4. **Variâncias são constantes (homogêneas) ao longo da linha de regressão.** Este pressuposto nos permite usar uma única constante σ^2 para a variância da linha de regressão. Se as variâncias dependem do *X*, podemos precisar de uma função de variância, ou uma família inteira de variâncias, atribuída a cada valor de *X* em particular. Variâncias não homogêneas são um problema comum em regressão e podem ser identificadas através de gráficos de diagnóstico (*ver*

[10] Apesar de pouco mencionada, fazemos um pressuposto similar de ausência de erro na variável categórica *X* na análise de variância (Capítulo 10). Na ANOVA, temos de assumir que os indivíduos são corretamente atribuídos aos grupos assinalados (p. ex., nenhum erro na identificação de espécies) e que todos os indivíduos dentro de um grupo receberam um tratamento idêntico (p. ex., todas as réplicas em um tratamento de "pH baixo" receberam o mesmo nível de pH).

próxima sessão). E de vez em quando é remediada através de transformações das variáveis X e Y originais (*ver* Capítulo 8). Contudo, não há nenhuma garantia de que uma transformação irá linearizar a relação e gerar variâncias homogêneas. Em tais casos, outros métodos (como a regressão não linear, descrita posteriormente neste capítulo) ou modelos lineares generalizados (McCullagh & Nelder, 1989) devem ser utilizados.

Se todos esses pressupostos forem atendidos, o método de mínimos quadrados provê **estimativas não tendenciosas** de todos os parâmetros do modelo. Eles não são tendenciosos, pois amostras repetidas da mesma população irão produzir estimativas de inclinações e interceptos que, em média, são as mesmas dos valores de inclinação e intercepto reais subjacentes à população.

TESTES DE DIAGNÓSTICO PARA A REGRESSÃO

Vimos até agora como obter estimativas por mínimos quadrados dos parâmetros de uma regressão linear, como testar hipóteses acerca desses parâmetros e a construir intervalos de confiança apropriados. Contudo, uma linha de regressão pode ser forçada através de qualquer conjunto de dados $[X, Y]$, independentemente de o modelo ser apropriado ou não. Nesta sessão, apresentaremos algumas ferramentas de diagnóstico para determinar quão bem a linha de regressão estimada se ajusta aos dados. Indiretamente, esses diagnósticos também ajudam a avaliar a extensão na qual os dados cumprem os pressupostos do modelo.

A mais importante ferramenta na análise de diagnósticos é o conjunto de resíduos, $\{d_i\}$, que representa as diferenças entre os valores observados (Y_i) e os valores preditos pelo modelo de regressão ($\hat{Y}_i$) da Equação 9.4. Os resíduos são usados para estimar a variância da regressão, além de fornecerem importantes informações acerca do ajuste do modelo aos dados.

Representação gráfica dos resíduos

Talvez o gráfico mais importante para a análise de diagnóstico de um modelo de regressão é o dos resíduos (d_i) *versus* os valores ajustados ($\hat{Y}_i$). Se o modelo linear se ajusta bem aos dados, o **gráfico dos resíduos** deve exibir uma dispersão de pontos que aproximadamente segue uma distribuição normal e são completamente não correlacionados com os valores ajustados (Figura 9.5A). No Capítulo 11, explicaremos o teste de Kolmogorov-Smirnov, que pode ser usado para comparar formalmente os resíduos a uma distribuição normal (*ver* Figura 11.4).

Dois tipos de problemas podem ser atacados com gráficos de resíduos. Primeiro, se os resíduos forem correlacionados com os valores ajustados significa que a relação não é linear. O modelo pode estar sistematicamente superestimando ou subestimando $\hat{Y}_i$ para valores altos da variável X (Figura 9.5B). Isso pode acontecer, por exemplo, quando uma linha reta é forçada através de dados que realmente representam uma relação assintótica, logarítmica ou outra rela-

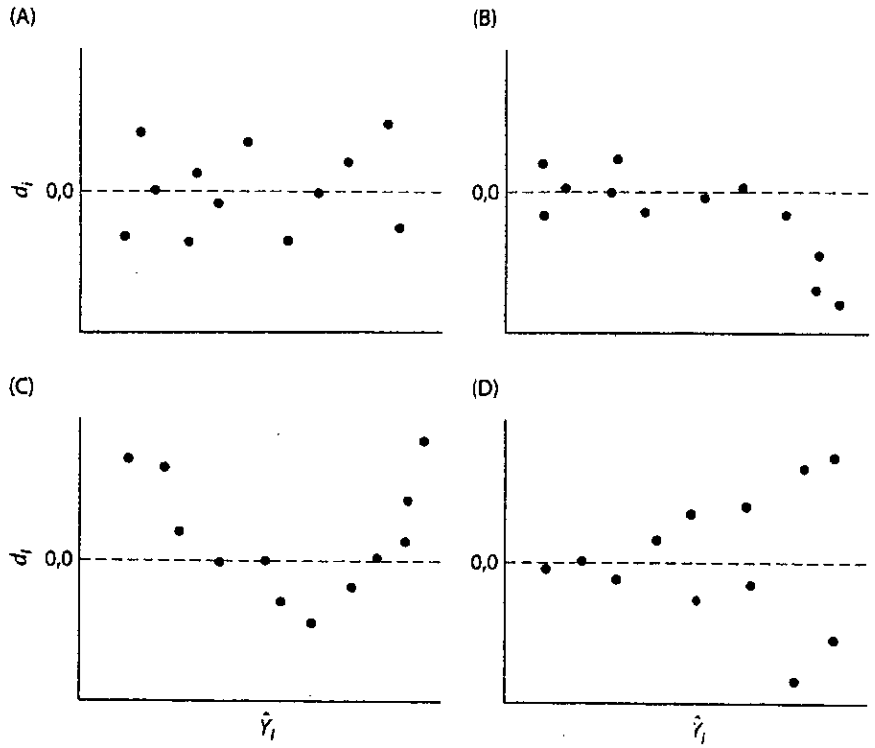


Figura 9.5 Padrões hipotéticos para gráficos de diagnóstico de resíduos (d_i), *versus* os valores ajustados ($\hat{Y}_i$) em regressão linear. (A) Distribuição esperada dos resíduos para um modelo linear com distribuição normal dos erros. Se os dados são bem ajustados pelo modelo linear, este é o padrão que deve ser encontrado nos resíduos. (B) Resíduos para um ajuste não linear. Neste caso, o modelo sistematicamente superestima os valores reais do *Y* conforme o *X* aumenta. Uma transformação matemática (p. ex., logaritmo, raiz quadrada ou inverso) pode produzir uma relação mais linear. (C) Resíduos para uma relação quadrática ou polinomial. Neste caso, os resíduos positivos grandes ocorrem para valores da variável *X*, que são muito pequenos, e para valores muito grandes. Uma transformação polinomial da variável *X* (X^2 ou alguma potência maior de *X*) pode produzir um ajuste linear. (D) Resíduos com heterocedasticidade (variância aumentando). Neste caso, os resíduos não são consistentemente negativos nem positivos, indicando que o ajuste do modelo é linear. Contudo, o tamanho médio dos resíduos aumenta com o *X* (heterocedasticidade), sugerindo que erros de medidas podem ser proporcionais ao tamanho da variável *X*. Uma transformação logarítmica ou da raiz quadrada pode corrigir esse problema. As transformações não são uma panaceia para as análises de regressão e nem sempre resultam em relações lineares.

ção não linear. Se os resíduos primeiro caem sobre os valores ajustados e depois caem abaixo, e então caem acima de novo, os dados podem indicar uma relação quadrática, em vez de linear (Figura 9.5C). Por fim, se os resíduos apresentam um funil de pontos crescente ou decrescente quando representados graficamente contra os valores ajustados (Figura 9.5D), a variância é **heterocedástica**: aumentando ou diminuindo com os valores ajustados. Transformações apropriadas dos dados, discutidas no Capítulo 8, podem atacar algum ou todos esses problemas. Os gráficos dos resíduos também podem destacar dados discrepantes

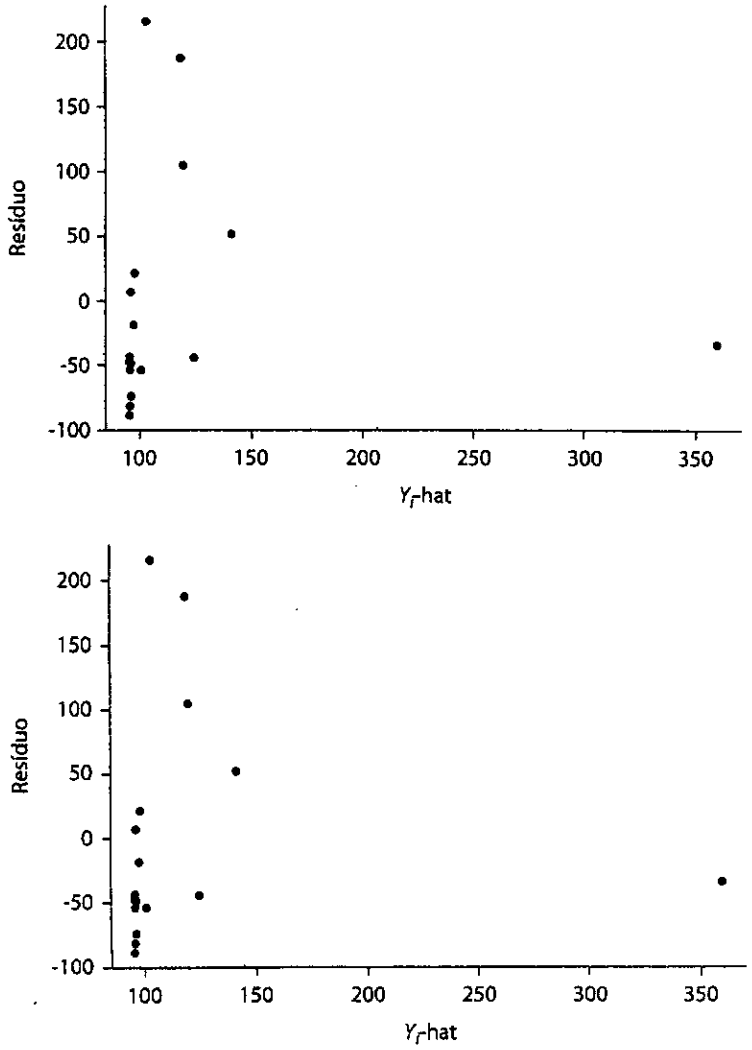


Figura 9.6 Gráfico dos resíduos para a relação espécie-área das plantas das ilhas de Galápagos (Tabela 8.2). Em um gráfico de resíduos, o eixo-*x* é o valor predito a partir da equação da regressão (Y_i) e o eixo-*y* é o resíduo, que é a diferença entre os valores observados e ajustados. Em um conjunto de dados correspondente às predições do modelo de regressão, o gráfico dos resíduos deve ser uma nuvem de pontos, com distribuição normal, centralizada no valor médio de 0 (*ver* Figura 9.5A). (A) Gráfico dos resíduos para a regressão com dados não transformados (Figura 8.5). Existem muitos resíduos negativos para valores ajustados pequenos, e a distribuição dos resíduos não parece normal. (B) Gráfico dos resíduos para a regressão com os mesmos dados após a transformação por $\log_{10}$ dos eixos-*X* e *Y* (Figuras 8.6 e 9.3). A transformação melhorou consideravelmente a distribuição dos resíduos, que não mais desviam de maneira sistemática nos valores grandes ou pequenos dos valores ajustados. O Capítulo 11 descreve o teste de Kolmogorov-Smirnov, usado para verificar se os resíduos se desviam de uma distribuição normal (*ver* Figura 11.4).

(*outliers*), pontos que caem muito distantes da predição da regressão do que a maioria dos outros dados.[11]

Para ver os efeitos das transformações, compare as Figuras 9.6A e 9.6B, as quais ilustram gráficos dos resíduos para os dados de Galápagos antes e após os dados serem transformados em $\log_{10}$. Sem a transformação, existem muitos resíduos negativos e todos agrupados acerca de valores muito pequenos de $\hat{Y}_i$ (Figura 9.6A). Após a transformação, os resíduos negativos e positivos são distribuídos de maneira uniforme e não associados com $\hat{Y}_i$ (Figura 9.6B).

Outros gráficos de diagnóstico

O gráfico dos resíduos pode ser feito não apenas contra os valores ajustados, mas também contra outras variáveis que podem ter sido medidas. A ideia é verificar a existência de qualquer variação escondida nos erros que possa ser atribuída a uma fonte sistemática. Por exemplo, poderíamos elaborar o gráfico dos resíduos da Figura 9.6B contra alguma medida de diversidade de hábitats em cada ilha. Se os resíduos forem correlacionados positivamente com a diversidade de hábitats, então as ilhas que tiverem mais espécies que o esperado, com base na área das ilhas, em geral terão maior diversidade de hábitats. De fato, esta fonte adicional de variação sistemática pode ser incluída em um modelo mais complexo de regressão múltipla, onde ajustamos coeficientes para duas ou mais variáveis preditoras. A representação gráfica dos resíduos contra variáveis preditoras geralmente é uma tática exploratória mais simples e confiável do que assumir um modelo com efeitos e interações lineares.

Também pode ser informativo colocar os resíduos contra a data ou a ordem da coleta. Esse gráfico pode indicar condições alteradas nas medições durante o período de coleta, como aumento na temperatura ao longo do dia em um estudo do comportamento de insetos, ou um medidor de pH com bateria fraca fornecendo resultados tendenciosos para as medidas que foram realizadas por último. Tais gráficos também nos lembram dos problemas inesperados que podem surgir ao não usar a randomização na coleta de dados. Se coletarmos todas as medidas de comportamento de uma espécie de inseto durante a manhã, ou medirmos o pH primeiro em todas as parcelas-controle, teremos introduzido uma inesperada fonte de variação de confusão em nossos dados.

A função de influência

Os gráficos dos resíduos fazem um bom trabalho para revelar a não linearidade, a heterocedasticidade e os dados discrepantes. Contudo, os dados com pontos in-

[11] Os **resíduos padronizados** (também chamados de **resíduos normalizados**) também são calculados por muitos pacotes de estatística. Os resíduos padronizados são escalonados para levar em conta a distância de cada ponto a partir do centro dos dados. Se não forem padronizados, dados bem distantes de $\bar{X}$ parecem ser relativamente bem ajustados pela regressão, enquanto pontos mais próximos do centro da nuvem de pontos parecem não se ajustar bem. Os resíduos padronizados também podem ser usados para testar se observações particulares são estatisticamente dados discrepantes (*outliers*) significantes. Outras medida de distância dos resíduos incluem a distância de Cook e os valores de alavancagem (*leverage*). *Ver* Sokal & Rohlf (1995) para detalhes.

fluenciáveis são potencialmente mais traiçoeiros. Esses dados podem não parecer discrepantes, mas eles têm uma influência irregular na estimativa da inclinação e do intercepto. Na pior das situações, dados com pontos influenciáveis que estão longe da nuvem típica de pontos podem "mudar" completamente a linha de regressão e dominar a estimativa da inclinação (Figura 7.3).

A melhor forma de detectar tais pontos é fazer um gráfico de **função de influência.** A ideia é simples e é, na verdade, uma forma de jackknife (*ver* Capítulo 5, nota de rodapé 2 e Capítulo 12). Pegue a primeira de n réplicas em seu conjunto de dados e a exclua da análise. Recalcule a inclinação, o intercepto e o valor de probabilidade. Agora, recoloque aquele dado de volta e exclua o segundo. Calcule novamente as estatísticas da regressão, substitua o dado e continue ao longo da lista. Se você tiver n pontos de dados originais, você terminará com n análises de regressão diferentes, cada uma com base em um total de $(n - 1)$ pontos. Agora, pegue as estimativas da inclinação e do intercepto de cada uma das análises e faça um único gráfico de pontos (Figura 9.7). Neste mesmo gráfico, coloque o ponto com os valores obtidos com o conjunto de dados completo. Esse gráfico das esti-

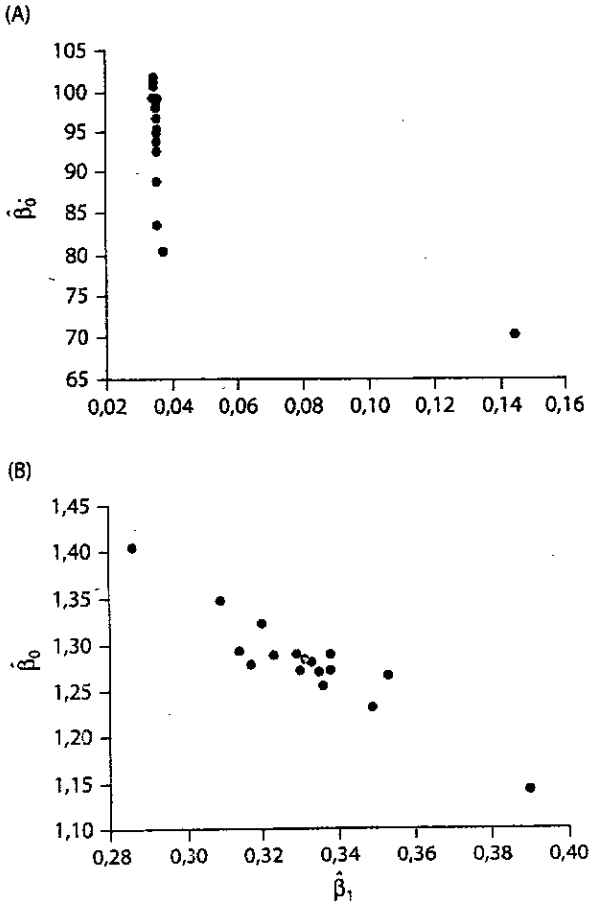


Figura 9.7 Função de influência para a regressão linear espécie-área dos dados de plantas das ilhas de Galápagos (Tabela 8.2). Em um gráfico de função de influência, cada ponto representa a inclinação ($\hat{\beta}_1$) e o intercepto ($\hat{\beta}_0$) recalculados após remover uma única observação do conjunto de dados. O ponto verde é a inclinação e o intercepto estimados com o conjunto de dados completo. (A) No modelo de regressão com dados não transformados, a informação da maior ilha, Albemarle (*ver* Figura 8.5), tem uma influência muito grande sobre a inclinação e o intercepto, gerando o ponto extremo no lado direito inferior do gráfico. (B) Função de influência calculada para a linha de regressão ajustada aos dados transformados em $\log_{10}$. Após a transformação logarítmica (Figuras 8.6 e 9.3), há, agora, uma nuvem de pontos mais homogênea na função de influência. Embora a inclinação e o intercepto mudem com a exclusão de cada observação, nenhuma tem uma influência extrema sobre o intercepto e a inclinação calculados. Neste conjunto de dados, a transformação logarítmica não melhora somente o ajuste linear dos dados, ela também estabiliza a estimativa da inclinação e do intercepto, de forma que não sejam dominadas por um ou dois pontos influenciáveis.

mativas do intercepto *versus* a da inclinação sempre terá uma inclinação negativa, pois, conforme a linha de regressão sobe, a inclinação aumenta e o intercepto diminui. Também pode ser informativo criar um histograma com os valores do r^2 ou as probabilidades de cauda, que são associadas com cada um daqueles modelos de regressão.

A função de influência ilustra quanto a estimativa dos parâmetros da regressão poderia mudar apenas com a exclusão de um único dado. Idealmente, a estimativa dos parâmetros com *jackknife* deveria agrupar fortemente ao redor da estimativa com o conjunto de dados completo. Um agrupamento de pontos pode sugerir que os valores de inclinação e intercepto são estáveis e não mudariam muito com a exclusão (ou adição) de um único dado. Por outro lado, se um dos pontos está muito distante do agrupamento, a inclinação e o intercepto são muito influenciados por aquele único dado – devemos ser cautelosos a respeito das conclusões acerca dessa análise. Similarmente, quando examinamos o histograma de valores de probabilidade, queremos ver todas as observações satisfatoriamente agrupadas ao redor do valor de *P* estimado com o conjunto de dados completo. Apesar de esperarmos por alguma variação, estaremos especialmente com problemas se uma das amostras de jackknife produzir um valor de *P* muito diferente de todo o resto. Em tal caso, rejeitar ou falhar em rejeitar a hipótese nula depende de uma única observação, o que é bastante perigoso.

Para os dados de Galápagos, a função de influência destaca a importância de usar uma transformação apropriada. Para os dados não transformados, a maior ilha do conjunto de dados domina a estimativa do β_1. Se este ponto for excluído, a inclinação aumenta de 0,035 para 0,114 (Figura 9.7A). Em contraste, a função de influência para os dados transformados mostra-se muito mais consistente, apesar de o β_1 ainda variar de 0,286 a 0,390 após a remoção de um único dado (Figura 9.7B). Para ambos os conjuntos de dados, os valores de probabilidade são estáveis e nunca estiveram acima de $P = 0{,}023$ para qualquer das análises de jackknife.

Usar o jackknife desta forma não é exclusividade da regressão. Qualquer análise estatística pode ser repetida com cada dado sistematicamente excluído. Isso pode consumir um pouco de tempo,[12] mas é uma excelente forma de avaliar a estabilidade e a validade geral de nossas conclusões.

MONTE CARLO E ANÁLISE BAYESIANA

A estimativa por mínimos quadrados e os testes de hipóteses são representativos da clássica análise paramétrica frequentista, pois assumem (em parte) que o termo do erro no modelo de regressão é uma variável aleatória normal, que os parâmetros têm valores fixos e reais a serem estimados e que os valores de *P* são derivados de distribuições de probabilidade fundamentadas em amostras infini-

[12] Na verdade, existem sábias soluções com matrizes para obter os valores de jackknife, apesar de não serem implementadas na maioria dos pacotes estatísticos. A função de influência é mais importante para pequenos conjuntos de dados que podem ser rapidamente analisados usando opções de seleção de casos, disponíveis na maioria dos *softwares* estatísticos.

tas. Como enfatizamos no Capítulo 5, as abordagens de Monte Carlo e bayesianas diferem filosoficamente ou em sua implementação (ou ambos), mas também podem ser usadas para análise de regressão.

Regressão linear usando métodos de Monte Carlo

Para a abordagem de Monte Carlo, ainda usamos todos os estimadores de mínimos quadrados para os parâmetros do modelo. Contudo, em uma análise de Monte Carlo, relaxamos o pressuposto de erros com distribuição normal, de forma que os testes de significância não sejam mais fundamentados em comparações de razões-F. Em vez disso, randomizamos, reamostramos os dados e simulamos os parâmetros diretamente para estimar sua distribuição.

Para uma análise de regressão linear, qual seria o método de randomização adequado? Podemos simplesmente aleatorizar os valores Y_i observados e pareá-los com os valores de X_i. Para tal conjunto de dados aleatorizado podemos calcular a inclinação, o intercepto, r^2, ou qualquer outra estatística de regressão usando os métodos descritos anteriormente. O método de randomização efetivamente quebra qualquer covariância entre as variáveis X e Y, e representa a hipótese nula de que a variável X (neste caso, a área da ilha) não tem nenhum efeito sobre a variável Y (neste caso, a riqueza de espécies). Intuitivamente, os valores de inclinação e do r^2 esperados devem ser aproximadamente zero para os conjuntos de dados aleatorizados, embora, por acaso, alguns tenham diferido. Para os dados de Galápagos, 5.000 randomizações dos dados geraram um conjunto de valores de inclinação com média de –0,002 (bem próximo de 0) e amplitude de –0,321 a 0,337 (Figura 9.8). Contudo, somente um dos valores simulados (0,337) excedeu a inclinação observada de 0,331. Portanto, a probabilidade em cauda estimada é 1/5.000 = 0,0002. Por comparação, a análise paramétrica forneceu essencialmente o mesmo resultado, com um valor de P igual a 0,0003 para $F_{1,15} = 21,118$ (Tabela 9.2).

Um resultado interessante da análise de Monte Carlo é que os valores de P para o teste do intercepto, da inclinação e do r^2 são todos idênticos. Naturalmente, a estimativa do parâmetro observada e a distribuição simulada para cada uma dessas variáveis são todas muito diferentes. Mas por que as probabilidades em cauda são as mesmas neste caso? A razão é que β_0, β_1 e r^2 não são algebricamente independentes uns dos outros, os valores aleatorizados refletem essa interdependência. Como consequência, a probabilidade em causa se torna a mesma para esses valores ajustados quando comparados a distribuições com base nos mesmos conjuntos de dados aleatorizados. Isto não acontece na análise paramétrica porque os valores de P não são determinados pela randomização dos dados.

Regressão linear usando métodos bayesianos

Nossa análise bayesiana começa com o pressuposto de que os parâmetros da equação de regressão não possuem valores reais a serem estimados, mas, em vez disso, são variáveis aleatórias (*ver* Capítulo 5). Portanto, a estimativa dos parâmetros de regressão – intercepto, inclinação e termo do erro – são reportados

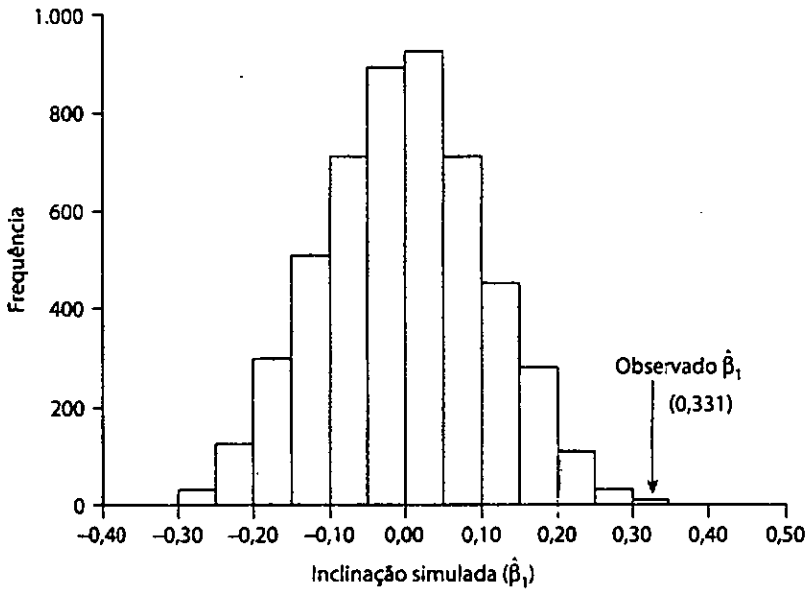


Figura 9.8 Análise de Monte Carlo dos valores de inclinação para a relação espécie-área com transformação log-log dos dados de espécies de plantas das ilhas de Galápagos (Tabela 8.2, *ver* Figuras 8.6 e 9.3). Cada observação no histograma representa a inclinação por mínimos quadrados de um conjunto de dados simulado em que os valores de *X* e *Y* foram aleatoriamente misturados. O histograma ilustra a distribuição de 5.000 desses conjuntos de dados. A inclinação observada com os dados originais é representada por uma seta no histograma. A inclinação observada excedeu todos, exceto um dos 5.000 valores simulados. Portanto, a probabilidade em causa estimada de obter um resultado tão extremo como este sob a hipótese nula é $P = 1/5.000 = 0,0002$. Essa probabilidade estimada é consistente com a calculada a partir da razão-F para o teste de regressão padrão desses dados ($P = 0,0003$; *ver* Tabela 9.2).

como distribuição de probabilidades, não como valores. Para executar qualquer análise bayesiana, incluindo a regressão, é necessário ter uma expectativa inicial para a forma dessas distribuições. Essas expectativas iniciais, chamadas de distribuição de probabilidades *a priori*, vêm de pesquisas anteriores ou do entendimento que temos do sistema de estudo. No caso de ausência de pesquisas anteriores, podemos usar distribuições de probabilidade de priores não informativas, como descrito no Capítulo 5. Contudo, no caso dos dados de Galápagos, podemos tirar vantagem do histórico de estudos sobre a relação entre a área das ilhas e a riqueza de espécies que antecedem a análise dos dados de Galapágos feita por Preston (1962).

Para tal exemplo, iremos executar a análise de regressão bayesiana usando dados que Preston (1962) compilou para outras relações espécie-área da literatura ecológica. Ele sumarizou a informação prévia disponível (em sua Tabela 8), e observou que, em outros 6 estudos, a inclinação média da relação espécie-área (gráfico $\log_{10}$-$\log_{10}$) = 0,278 com desvio-padrão de 0,036. Preston (1962) não divulgou sua estimativa do intercepto, mas nós a calculamos a partir dos dados que ele divulgou: o intercepto médio = 0,854 e o desvio-padrão = 0,091). Por fim,

precisamos de uma estimativa da variância do erro da regressão ε, que calculamos a partir de dados, por ele fornecidos, como 0,234.

Com essas informações *a priori,* usamos o Teorema de Bayes para calcular uma nova distribuição (a **distribuição de probabilidades *a posteriori***) de cada um desses parâmetros. As distribuições de probabilidade *a posteriori* levam em consideração a informação prévia mais os dados das ilhas de Galápagos. Os resultados dessas análises são:

$$\hat{\beta}_0 \sim N(1{,}863;\ 0{,}157)$$
$$\hat{\beta}_1 \sim N(0{,}328;\ 0{,}145)$$
$$\sigma^2 = 0{,}420$$

Em palavras, o intercepto é uma variável aleatória normal com média = 1,863 e desvio-padrão = 0,157, a inclinação é uma variável aleatória normal com média = 0,328 e desvio-padrão = 0,145, e a variância da regressão é uma variável aleatória normal com média 0 e variância 0,420.[13]

Como interpretamos esses resultados? A primeira coisa é re-enfatizar que, em uma análise bayesiana, os parâmetros são considerados como variáveis aleatórias – não têm valores fixos. Assim, o modelo linear (Equação 9.3) precisa ser re-expresso como:

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \qquad (9.32)$$

Em palavras, cada observação Y_i é tirada de uma distribuição normal com média = $\beta_0 + \beta_1 X_i$ e variância = σ^2. Similarmente, tanto o intercepto β_0 quanto a inclinação β_1 são, por si só, variáveis aleatórias normais com médias estimadas $\hat{\beta}_0$ e $\hat{\beta}_1$, e variâncias estimadas correspondentes s_0^2 e s_1^2.

Como essas estimativas se aplicam a nossos dados? Primeiro, a linha de regressão tem uma inclinação de 0,328 (o valor esperado para β_1) e um intercepto de 1,863. A inclinação difere somente um pouco da linha de regressão por mínimos quadrados, conforme mostrado na Figura 9.4, que tem inclinação de 0,331. O intercepto é substancialmente maior do que no ajuste por mínimos quadrados, sugerindo que um número maior de ilhas pequenas contribuiu mais para essa regressão que as poucas ilhas grandes.

Segundo, podemos produzir um intervalo de credibilidade de 95%. A computação é essencialmente a mesma mostrada nas Equações 9.25 e 9.27, mas a interpretação do intervalo de credibilidade bayesiano é que o valor médio poderia cair dentro do intervalo de credibilidade 95% das vezes. Isso difere da interpretação do

[13] Alternativamente poderíamos conduzir essa análise sem nenhuma informação *a priori*, como foi discutido no Capítulo 4. Esta análise bayesiana mais "objetiva" usa distribuições de probabilidade de priores não informativas sobre os parâmetros da regressão. Os resultados para as análises usando priores não informativas são ($\hat{\beta}_0 \sim N(1{,}866;\ 0{,}084)$, $\hat{\beta}_1 \sim N(0{,}329;\ 0{,}077)$ e $\sigma^2 = 0{,}119$). A falta de diferenças substanciais entre esses resultados e aqueles com base na distribuição de priores informadas ilustra que bons dados superam a opinião subjetiva.

intervalo de confiança de 95%, que é de que o intervalo de confiança irá incluir o valor real 95% das vezes (*ver* Capítulo 3).

Terceiro, se precisarmos predizer a riqueza de espécies para uma dada ilha de tamanho X_i, o número de espécies *predito* poderia ser encontrado primeiro tirando um valor de β_1 de uma distribuição normal com média = 0,328 e desvio-padrão = 0,145, e então multiplicando o valor por X_i. Depois, tiramos um valor para o intercepto β_0 de uma distribuição normal com média = 1,863 e desvio-padrão = 0,157. Por fim, adicionamos um "ruído" a essa predição, adicionando um termo do erro de uma distribuição normal com $X = 0$ e $\sigma^2 = 0,42$.

Por fim, algumas palavras acerca do teste de hipóteses. Não há uma forma para calcular um valor de *P* em uma análise bayesiana. Não há um valor de *P*, pois não estamos testando uma hipótese acerca de quão improváveis nossos dados são, dadas as hipóteses [como na estimativa frequentista de $P(\text{dados}|H_0)$] ou quão frequentemente podemos esperar obter esses resultados dado apenas em razão do acaso (como na análise de Monte Carlo). Em vez disso, estimamos valores para os parâmetros. As análises paramétricas e de Monte Carlo também fazem o mesmo, mas o objetivo primário na maioria das vezes é determinar se a estimativa é significativamente diferente de zero. As análises frequentistas também podem incorporar informações *a priori*, porém ainda no contexto de um teste de hipóteses binário (sim/não). Por exemplo, se tivermos informação prévia de que a inclinação da relação espécie-área é $\beta_1 = 0,26$, e nossos dados de Galápagos sugiram que seja $\beta_1 = 0,033$, podemos testar se estes são "significativamente diferentes" dos dados previamente publicados. Se a resposta for "sim", o que você conclui? E se a resposta for "não"? Uma análise bayesiana assume que você está interessado em usar todos os dados que tem para estimar a inclinação e o intercepto da relação espécie-área.

OUTROS TIPOS DE ANÁLISES DE REGRESSÃO

O modelo básico de regressão que desenvolvemos apenas pincela a superfície de tipos de análises que podem ser feitas com variáveis contínuas preditoras ou respostas. Faremos apenas uma breve descrição dos outros tipos de análises de regressão. Livros inteiros se dedicam a cada um desses tópicos individualmente, portanto não pretendemos fornecer nada além de uma breve introdução. Não obstante, muitos dos mesmos pressupostos, restrições e problemas que descrevemos para a regressão linear simples também se aplicam para esses métodos.

Regressão robusta

O modelo de regressão linear estima a inclinação, o intercepto e a variância, minimizando a soma dos quadrados dos resíduos (*SQR*) (*ver* Equação 9.5). Se os erros seguirem uma distribuição normal, a soma dos quadrados dos resíduos fornecerá uma estimativa sem viés dos parâmetros do modelo. As estimativas por mínimos quadrados são sensíveis a dados discrepantes (*outliers*), porque elas dão maior peso aos resíduos maiores. Por exemplo, um resíduo de 2 contribui com $2^2 = 4$ para a *SQR*, mas um resíduo de 3 contribui com $3^2 = 9$ à *SQR*. A penalidade adi-

cionada a um resíduo grande é apropriada, pois esses valores maiores serão relativamente raros se os erros amostrados são retirados de uma distribuição normal.

Contudo, quando dados discrepantes verdadeiros estiverem presentes – dados aberrantes (incluindo os errôneos) que não foram amostrados a partir da mesma distribuição –, podem seriamente inflar a estimativa da variância. O Capítulo 8 discutiu vários métodos para identificar e tratar dados discrepantes. Outra estratégia é ajustar o modelo usando uma função residual, que não seja sensível à presença de dados discrepantes. As técnicas de **regressão robusta** usam diferentes funções matemáticas para quantificar a variação residual. Por exemplo, em vez de elevar cada resíduo ao quadrado, poderíamos usar seu valor absoluto:

$$\text{resíduo} = \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right| = \sum_{i=1}^{n} |d_i| \qquad (9.33)$$

Como na SQR, este *resíduo* é maior se os desvios individuais (d_i) forem grandes. Contudo, os muito grandes não são penalizados com mais peso porque seus valores não são elevados ao quadrado. Esse peso é apropriado se você acredita que os dados foram retirados de uma distribuição que tem caudas relativamente pesadas (leptocúrticas; *ver* Capítulo 3) se comparadas a uma distribuição normal. Outras medidas que dão mais ou menos peso aos grandes desvios podem ser usadas.

Entretanto, uma vez que abandonamos a SQR, não podemos mais usar as fórmulas simples para obter estimativas dos parâmetros da regressão. Em vez disso, técnicas de computador iterativas são necessárias para encontrar a combinação de parâmetros que minimizam os resíduos (*ver* Nota de rodapé 5, Capítulo 4). Para ilustrar a regressão robusta, a Figura 9.9 mostra os dados de Galápagos com um dado discrepante adicional. Uma linha de regressão padrão para esse novo conjunto de dados dá uma inclinação de 0,241, 27% mais rasa que a inclinação real de 0,331, mostrada na Figura 9.4. Duas diferentes técnicas de regressão robusta, usando **mínimos quadrados aparados** (*least-trimmed squares*) (Davies, 1993) e **estimadores-M** (Huber, 1981), foram aplicadas aos dados mostrados na Figura 9.9. Como na regressão linear padrão, esses métodos de regressão robusta assumem que a variável X é medida sem erro.

A regressão por mínimos quadrados aparados minimiza a soma dos quadrados dos resíduos (Equação 9.5) aparando uma porcentagem das observações extremas. Para um corte de 10% – removendo o decil superior da soma dos quadrados dos resíduos –, a inclinação predita para a relação espécie-área para os dados ilustrados na Figura 9.9 é 0,283, uma melhora de 17% sobre a regressão linear básica. O intercepto para a regressão de mínimos quadrados aparados é 1.401, algo maior que o intercepto real.

Os estimadores-M minimizam o resíduo:

$$\text{resíduo} = \sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i b}{s}\right) + n \log s \qquad (9.34)$$

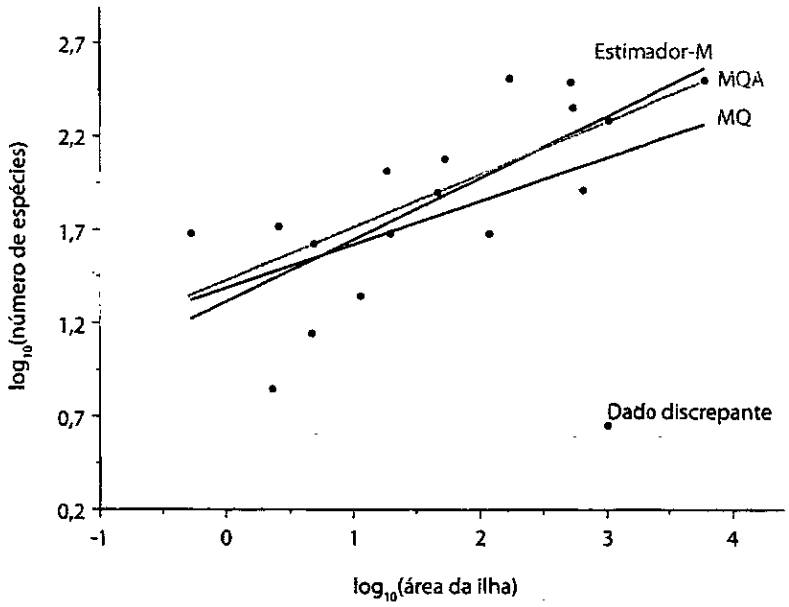


Figura 9.9 Três regressões para a relação espécie-área de espécies de plantas das ilhas de Galápagos (Tabela 8.2; *ver* Figuras 8.6 e 9.3) com a adição de um dado discrepante artificial (*outlier*). A regressão linear padrão (MQ) é relativamente sensível a dados discrepantes. Enquanto a estimativa da inclinação original foi de 0,331, a nova inclinação é de apenas 0,233, e foi arrastada para baixo pelo dado discrepante. O método de mínimos quadrados aparados (MQA) descarta 10% dos dados extremos (5% da parte de cima e 5% da parte de baixo). A inclinação para a regressão por MQA é 0,283 – um tanto mais próximo à estimativa original, de 0,331, embora a estimativa do intercepto tenha aumentado para 1,432. A estimativa-M recupera a estimativa "correta" da inclinação (0,331) e do intercepto (1,319), embora a estimativa da variância seja praticamente o dobro daquela da regressão linear que não incluía o dado discrepante. Esses métodos de regressão robusta são úteis para ajustar equações de regressão a dados muito variáveis, incluindo dados discrepantes.

onde X_i e Y_i são valores das variáveis preditora e resposta, respectivamente; $\rho = -\log f$, onde f é uma função de densidade de probabilidade dos resíduos, ε, que são escalonados por um peso s, $[f(\varepsilon/s)]/s$, e b é uma estimativa da inclinação β_1. O valor de b que minimiza a Equação 9.34 para um dado valor de s é um estimador-M robusto da inclinação da linha de regressão β_1. O estimador-M efetivamente pondera os dados pelos seus resíduos, de forma que os com maiores resíduos contribuam menos para a estimativa da inclinação.

Naturalmente, é necessário fornecer uma estimativa do s para resolver a equação. Venables & Ripley (2002) sugerem que o s pode ser estimado resolvendo a equação:

$$\sum_{i=1}^{n} \Psi\left(\frac{Y_i - X_i b}{s}\right)^2 = n \qquad (9.35)$$

onde ψ é a derivada do ρ usado na Equação 9.34. Felizmente, todos esses cálculos podem ser feitos com facilidade em programas como o S-Plus ou o R, que possuem rotinas disponíveis para regressão robusta (Venables & Ripley, 2002).

A regressão robusta por estimador-M produz a inclinação correta de 0,331, mas a variância em torno dessa estimativa é igual a 0,25 – mais de três vezes maior que a variância (0,07), sem o dado discrepante. Da mesma forma, o intercepto é correto em 1,319, mas sua variância é quase o dobro (0,71 *versus* 0,43). Por fim, a estimativa geral da variância do erro da regressão é igual a 0,125, mais de 30% superior à da regressão simples feita com os dados originais.

Uma estratégia de compromisso sensato é usar os métodos comuns de mínimos quadrados para testar hipóteses sobre a significância da inclinação e do intercepto. Com a presença de dados discrepantes, esses testes serão relativamente conservativos, pois os dados irão inflar a estimativa da variância. Contudo, a regressão robusta poderia ser usada para construir intervalos de predição. A única advertência é que os erros são, na verdade, tirados de uma distribuição normal com variância grande, a regressão robusta irá subestimar a variância e você ocasionalmente poderá ter uma surpresa caso faça predições a partir do modelo de regressão robusta.

Regressão quantílica

A regressão linear simples ajusta a linha através do centro da nuvem de pontos e é apropriada para descrever uma relação direta de causa e efeito entre as variáveis *X* e *Y*. No entanto, se a variável *X* atua como fator limitante, ela pode impor um teto para os valores de *Y*. Assim, a variável *X* pode controlar o valor máximo para o *Y*, mas não ter nenhum efeito abaixo desse máximo. O resultado poderia ser um gráfico com forma de triângulo. Por exemplo, a Figura 9.10 retrata a relação entre a biomassa anual de nozes do carvalho e um "índice de adequação das nozes" medido em 43 parcelas amostrais de 0,2 hectares em Missouri – EUA (Schroeder & Vangilder, 1997). Os dados sugerem que a adequação baixa restringe a biomassa anual das nozes, porém, com adequação alta, a biomassa das nozes em uma parcela pode ser alta ou baixa, dependendo de outros fatores limitantes.

A **regressão quantílica** minimiza os desvios a partir da linha de regressão, mas a função de minimização é assimétrica: desvios positivos e negativos são ponderados diferentemente:

$$\text{resíduo} = \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right| h \qquad (9.36)$$

Como na regressão robusta, a função minimiza o valor absoluto dos desvios, de forma que não seja sensível aos dados discrepantes. No entanto, a característica-chave é o multiplicador *h*. O valor do *h* é o quantil a ser estimado. Se o sinal do desvio dentro do valor absoluto for positivo, ele será multiplicado por *h*. Se o desvio for negativo, ele será multiplicado por $(1,0 - h)$. Essa minimização assimé-

trica ajusta uma linha de regressão através da região superior dos dados para um h grande e através da região inferior para um h pequeno. Se $h = 0,5$, a linha de regressão passa através do centro da nuvem de dados e é equivalente à regressão robusta, usando a Equação 9.5.

O resultado é uma família de linhas de regressão que caracterizam os limites superiores e inferiores do conjunto de dados (Figura 9.10). Se a variável X for o único fator que atinge a Y, essas regressões quantílicas serão grosseiramente paralelas umas às outras. Contudo, se outras variáveis estão em jogo, a inclinação das regressões quantílicas superiores serão bem mais íngremes que a linha de regressão padrão. Esse padrão pode ser uma indicativa de um limite superior ou de um fator limitante em ação (Cade & Noon, 2003).

Alguns cuidados são necessários com a regressão quantílica. A primeira é que a escolha de qual quantil usar é um tanto arbitrária. Além disso, quanto mais extremo for o quantil, menor o tamanho amostral, limitando o poder do teste. Ainda, as regressões quantílicas serão dominadas por dados discrepantes, mesmo

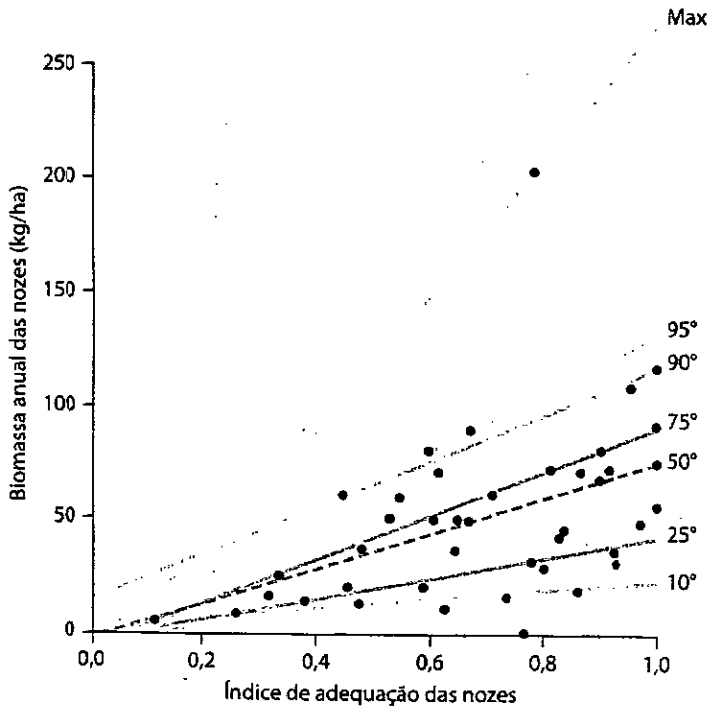


Figura 9.10 Ilustração da regressão quantílica. A variável X é uma medida de adequação das nozes em parcelas de floresta, com base em características das florestas de carvalho para $n = 43$ parcelas amostrais de 0,2 ha em Missouri (dados de Schroeder & Vangilder, 1997). A variável Y é a biomassa anual das nozes em cada parcela. As linhas sólidas correspondem às inclinações da regressão quantílica. A linha pontilhada (= 50-ésimo percentil) é a linha da regressão padrão por mínimos quadrados que passa através do centro da nuvem de pontos. A regressão quantílica é apropriada quando há um fator limitante que estabelece um teto para a variável resposta. (Adaptação e simplificação da Figura 3, de Cade et al., 1999.)

sabendo que a Equação 9.36 minimiza seus efeitos relativos à solução por mínimos quadrados. De fato, a regressão quantílica, por definição, é uma linha que passa através dos dados extremos, motivo pelo qual você precisa estar certo de que esses valores extremos não representem apenas erros.

Por fim, a regressão quantílica pode não ser necessária, a menos que a hipótese realmente seja sobre a existência de um teto, ou de um fator limitante. Com frequência, a regressão quantílica é usada com dados que produzem gráficos "triangulares". Contudo, esses triângulos podem simplesmente refletir heterocedasticidade – uma variância que aumenta com valores altos da variável X. A transformação dos dados e a análise cuidadosa dos resíduos podem revelar que a regressão linear simples é mais apropriada. *Ver* Thompson et al. (1996), Garvey et al. (1998), Scharf et al. (1998) e Cade et al. (1999) para outros métodos de regressão usando dados ecológicos bivariados.

Regressão logística

A **regressão logística** é um caso especial em que a variável Y é categórica, e não contínua. O caso mais simples é de uma variável Y dicotômica. Por exemplo, conduzimos censos temporais em 42 folhas de uma planta carnívora (*Darlingtonia californica*) escolhidas de maneira aleatória, uma planta-jarro que captura insetos (*ver* Nota de rodapé 11, Capítulo 8). Registramos visitas de vespas a 10 das 42 folhas (Figura 9.11). Podemos usar a regressão logística para testar a hipótese na qual a probabilidade de visitação é relacionada à altura das folhas.[14]

Você pode ficar tentado a forçar uma linha de regressão através desses dados, mas mesmo se eles fossem ordenados com perfeição, a relação pode não ser linear. Em vez disso, a melhor curva de ajuste tem a forma de S, ou logística, aumentando a partir de um mínimo até uma assíntota máxima. Esse tipo de curva pode ser descrita por uma função com dois parâmetros, β_0 e β_1:

$$p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \qquad (9.37)$$

O parâmetro β_0 é um tipo de intercepto, pois ele determina a probabilidade de sucesso ($Y_i = 1$), p, quando $X = 0$, então $p = 0{,}5$. O parâmetro β_1 é similar ao de inclinação, pois determina quão íngreme é o aumento da curva até o valor máximo de $p = 1{,}0$. Juntos, β_0 e β_1 especificam a amplitude da variável X em que a maior parte do aumento ocorre, além de determinar quão rapidamente o valor da probabilidade aumenta de 0,0 a 1,0.

[14] É claro que poderíamos usar os bem mais simples teste-t ou ANOVA (Capítulo 10) para comparar a altura da plantas visitadas *versus* não visitadas. Contudo, uma ANOVA com esses dados subitamente torna a causa-e-efeito o foco da análise: a hipótese é de que a altura das plantas influencia a probabilidade de captura, que é o que está sendo testado com a regressão logística. A estrutura da ANOVA implica que a variável categórica *visitação* de alguma forma causa ou é responsável pela diferença na variável resposta contínua, altura das folhas.

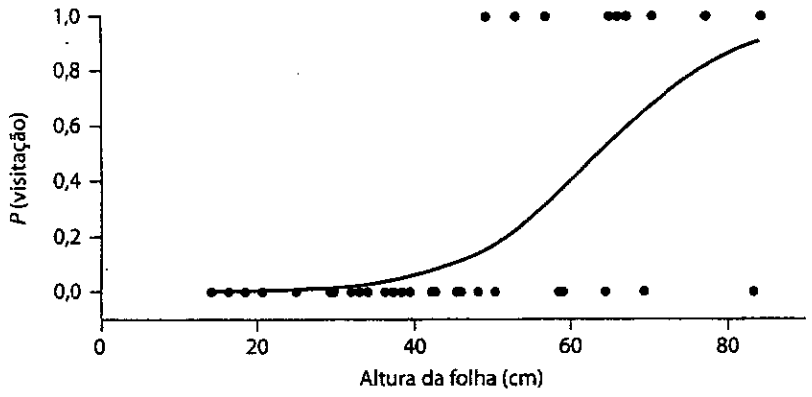


Figura 9.11 Relação entre a altura das folhas e a visitação de vespas à planta carnívora *Darlingtonia californica*. Cada ponto representa uma planta diferente em uma população das Montanhas Siskiyou no sudeste de Oregon (Ellison & Gotelli, dados não publicados). O eixo *X* representa a altura das folhas, uma variável preditora contínua. O eixo *Y* é a probabilidade de visitação de vespas. Apesar de ser uma variável contínua, na verdade os dados são discretos, pois a planta é visitada (1) ou não (0). A regressão logística ajusta uma curva em forma de S (= logística) a esses dados. A regressão logística é usada aqui porque a variável resposta é discreta, de forma que a relação tem uma assíntota superior e uma inferior. O modelo é ajustado usando a transformação logit (Equação 9.38). Os parâmetros com melhor ajuste (usando máxima verossimilhança por ajuste iterativo) são $\hat{\beta}_0 = -7{,}293$ e $\hat{\beta}_1 = 0{,}115$. O valor de *P* para um teste da hipótese nula de que $\beta_1 = 0$ é 0,002, sugerindo que a probabilidade de visitação de vespas aumenta conforme a altura das folhas.

A razão para usar a Equação 9.37 é que, com um pouco mais de álgebra, podemos transformá-la como a seguir:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \qquad (9.38)$$

Essa transformação da variável *Y*, chamada de **transformação logit**, converte a curva logística em forma de S em uma linha reta. Embora essa transformação de fato seja linear para a variável *X*, não podemos aplicá-la diretamente aos nossos dados. Se eles consistem somente em 0's e 1's para plantas de diferentes tamanhos, a Equação 9.38 não pode ser resolvida, pois $\ln[p/(1-p)]$ é indefinido para $p = 1$ ou $p = 0$. Porém, mesmo se os dados consistissem de estimativas de p com base em múltiplas observações de plantas com o mesmo tamanho, ainda não seria apropriado usar os estimadores de mínimos quadrados, pois o termo do erro segue uma distribuição binomial, em vez de uma distribuição normal.

Além disso, usamos a abordagem de máxima verossimilhança, que fornece estimativa dos parâmetros que tornam os dados mais prováveis (*ver* Nota de rodapé 13, Capítulo 5). Essa solução inclui uma estimativa da variância do erro da regressão, que pode ser usada para testar a hipótese nula acerca dos valores dos parâmetros e para construir intervalos de confiança, como na regressão padrão.

Para os dados da *Darlingtonia*, as estimativas de máxima verossimilhança dos parâmetros são $\hat{\beta}_0 = -7{,}293$ e $\hat{\beta}_1 = 0{,}115$. O teste da hipótese nula de que $\beta_1 = 0$ gera um valor de $P = 0{,}002$ sugere que a probabilidade de visitação das vespas aumenta com o tamanho das folhas.

Regressão não linear

Embora o método de mínimos quadrados descreva relações lineares entre variáveis, ele pode ser adaptado para funções não lineares. Por exemplo, a dupla transformação logarítmica converte uma função de potência para as variáveis X e Y, ($Y = aX^b$), em uma relação linear entre $\log(X)$ e $\log(Y)$, [$\log(Y) = \log(a) + b \times \log(X)$]. Contudo, nem todas as funções podem ser transformadas assim. Por exemplo, muitas funções não lineares foram propostas para descrever a resposta funcional de predadores – ou seja, a mudança na taxa de alimentação dos predadores em função da densidade de presas. Se um predador forrageia aleatoriamente sobre um recurso de presas que é exaurido com o passar do tempo, a relação entre o número consumido (N_c) e o inicial (N_0) é (Rogers, 1972):

$$N_c = N_0\left(1 - e^{a(T_h N_c - T)}\right) \qquad (9.39)$$

Nesta equação existem três parâmetros a serem estimados: a, a taxa de ataque; T_h, o tempo de manuseio por presa; e T, o tempo total de manuseio de todas as presas. A variável Y é N_c, o número consumido, e a variável X é N_0, o número inicial de presas.

Não há uma transformação algébrica que lineariza a Equação 9.39, de modo que o método de mínimos quadrados não pode ser usado. Em vez disso, uma **regressão não linear** é usada para ajustar os parâmetros do modelo da função sem transformação. Como na regressão logística (um tipo particular de regressão não linear), métodos iterativos são usados para gerar parâmetros que minimizam os desvios de mínimos quadrados e permitir os testes de hipóteses e cálculo de intervalos de confiança.

Mesmo quando uma transformação produz um modelo linear, a análise de mínimos quadrados assume que os termos do erro, ε_i, *dos dados transformados* possuem distribuição normal. Porém, se na função original os termos do erro têm distribuição normal, a transformação não irá preservar a normalidade. Em tais casos, a regressão linear também será necessária. Trexler & Travis (1993) e Juliano (2001) fornecem boas introduções a análises ecológicas usando regressões não lineares.

Regressão múltipla

O modelo de regressão linear com uma única variável preditora X pode ser facilmente estendido a duas ou mais variáveis preditoras, ou a polinômios de ordens maiores de uma única variável preditora. Por exemplo, suponha que suspeitamos que a riqueza de espécies atinge o máximo em ilhas com tamanho intermediário,

talvez por causa de gradientes na frequência ou intensidade de distúrbios. Poderíamos ajustar um polinômio de segunda ordem que inclua um termo quadrado para área da ilha:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \qquad (9.40)$$

Esta equação descreve uma função com um pico na riqueza de espécies em um tamanho de ilhas intermediário.[15] A Equação 9.40 é um exemplo de **regressão múltipla**, pois agora há duas variáveis preditoras, X e X^2, que contribuem para Y. Entretanto, ele ainda é considerado um modelo de regressão linear (embora seja do tipo múltipla), pois os parâmetros β_i na Equação 9.40 são obtidos por meio de equações lineares. Se os dados forem modelados como uma função linear simples, ignoramos o termo polinomial, e o componente de variação sistemática X^2 será incorretamente unido ao termo do erro:

$$\varepsilon_i' = \beta_2 X_i^2 + \varepsilon_i \qquad (9.41)$$

Um exemplo mais familiar de regressão múltipla é quando duas ou mais variáveis preditoras são medidas em cada réplica. Por exemplo, em um estudo da variação na riqueza de espécies de formigas (S) em florestas de pântanos da Nova Inglaterra (Gotelli & Ellison, 2002a,b), medimos a latitude e a altitude de cada sítio amostral e usamos as duas variáveis em uma equação de regressão múltipla:

$$\log_{10}(S\ floresta) = 4{,}879 - 0{,}089(latitude) - 0{,}001(altitude) \qquad (9.42)$$

Ambos os parâmetros de inclinação são negativos porque a riqueza de espécies declina em maiores altitudes e maiores latitudes. Neste modelo temos os chamados **parâmetros parciais da regressão**, pois a soma dos quadrados dos resíduos de outras variáveis já foram estatisticamente consideradas. Por exemplo, o parâmetro para a latitude poderia ser encontrado primeiro fazendo a regressão da riqueza de espécies pela altitude, e depois dos resíduos contra a latitude. Inversamente, o parâmetro para a altitude pode ser obtido fazendo a regressão dos resíduos, da regressão riqueza-latitude, contra a altitude.

Como os parâmetros parciais da regressão fundamentam-se na variação residual não explicada pelas outras variáveis, eles frequentemente não serão equiva-

[15] Esse pico pode ser encontrado através da derivada da Equação 9.40 e em seu ajuste para ser igual a 0. Assim, a riqueza de espécies máxima ocorre em $X = \beta_1/\beta_2$. Embora esta equação produza uma curva não linear para o gráfico de Y *versus* X, o modelo ainda será uma soma linear de forma $\Sigma\beta_i X_i$, onde X_i é a variável preditora medida (que pode ser uma variável transformada) e os β_i são os parâmetros da regressão ajustados.

lentes aos estimados em modelos de regressão simples. Por exemplo, se fizermos a regressão da riqueza de espécies somente pala latitude, o resultado é:

$$\log_{10}(S\,floresta) = 5{,}447 - 0{,}105(latitude) \qquad (9.43)$$

E, se fizermos a regressão da riqueza de espécies somente pela altitude, temos:

$$\log_{10}(S\,floresta) = 1{,}087 - 0{,}001(altitude) \qquad (9.44)$$

Com exceção do termo da altitude, todos os parâmetros diferem entre os modelos de regressão linear simples (Equações 9.43 e 9.44) e dos da regressão múltipla (Equação 9.42).

Na regressão linear com uma única variável preditora, a função pode ser colocada no gráfico como uma linha em duas dimensões (Figura 9.12). Com duas variáveis preditoras, a equação da regressão múltipla pode ser colocada no gráfico como um plano em um espaço de coordenadas tridimensionais (Figura 9.13). O "piso" do espaço representa os dois eixos para as duas variáveis preditoras, e a dimensão vertical representa a variável resposta. Cada réplica consiste em três medidas (variáveis Y, X_1 e X_2), de forma que os dados podem ser plotados como uma nuvem de pontos no espaço tridimensional. A solução por mínimos quadrados traça um plano através da nuvem de pontos. O plano é posicionado de forma

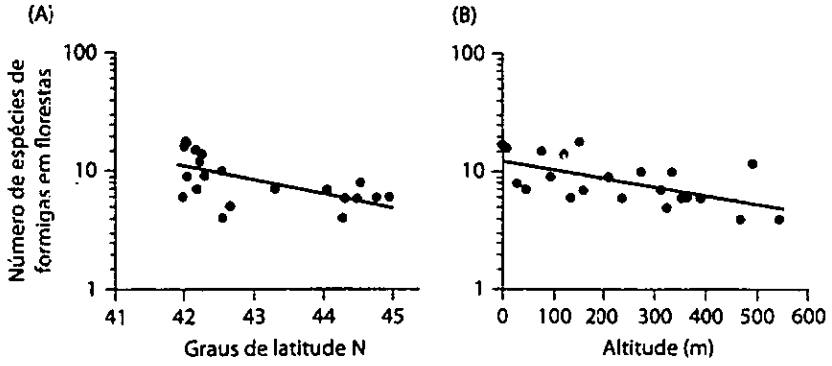


Figura 9.12 Densidade de formigas de floresta em função de (A) latitude e (B) altitude. Cada ponto ($n = 22$) é o número de espécies de formiga registrado em parcelas de floresta com 64 m^2 na Nova Inglaterra (Vermont, Massachusetts e Connecticut). A linha de regressão por mínimos quadrados é mostrada para cada variável. Note a transformação por $\log_{10}$ do eixo Y. Para a regressão pela latitude, a equação é $\log_{10}$(número de espécies de formigas) = 5,447 – 0,105 × (latitude); $r^2 = 0{,}334$. Para a regressão pela altitude, a equação é $\log_{10}$(número de espécies de formigas) = 1,087 – 0,001 × (altitude), $r^2 = 0{,}353$. As inclinações negativas indicam que a riqueza de espécies declina em maiores latitudes e altitudes. (Detalhes sobre os dados e amostragem em Gotelli & Ellison, 2002a,b.)

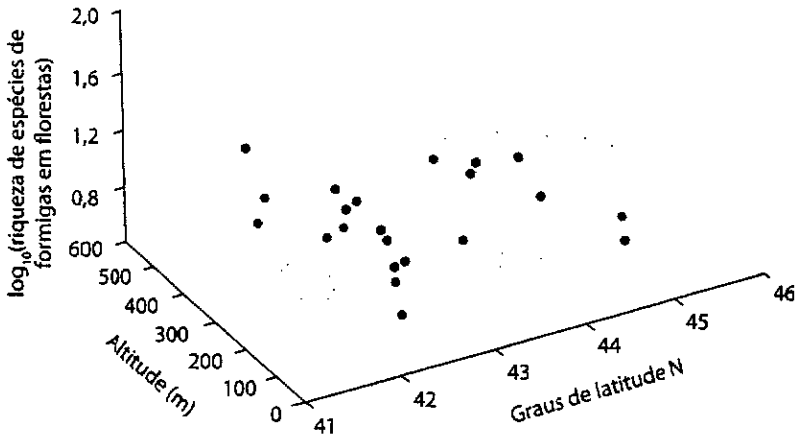


Figura 9.13 Representação tridimensional dos dados de uma regressão múltipla. A solução de mínimos quadrados é um plano que passa através da nuvem de dados. Cada ponto ($n = 22$) é o número de espécies de formigas registrado em parcelas de floresta com 64 m^2 no norte da Nova Inglaterra (Vermont, Massachusetts e Connecticut). A variável *X* representa a latitude (a primeira variável preditora), a *Y* representa a altitude (a segunda variável preditora) e a *Z*, o $\log_{10}$ do número de espécies de formigas registrado (a variável resposta). Uma equação de regressão múltipla foi ajustada à equação: $\log_{10}$(número de espécies de formigas) = 4,879 – 0,0089 × (latitude) – 0,001 × (altitude), $r^2 = 0,583$. Note que o valor do r^2, que é uma medida do ajuste dos dados ao modelo, é maior para o de regressão múltipla que para ambos os modelos de regressão simples com base em apenas uma das variáveis preditoras (Figura 9.12). A solução para uma regressão linear com uma variável preditora é uma reta, enquanto a solução para uma regressão múltipla com duas variáveis preditoras é um plano, mostrado nessa representação tridimensional. Os resíduos são calculados como a distância vertical de cada dado até o plano predito. (Detalhes sobre os dados e a amostragem em Gotelli & Ellison, 2002a,b.)

que a soma dos quadrados dos desvios verticais de todos os pontos até o plano seja minimizada.

As soluções por matrizes para a regressão múltipla são descritas no Apêndice, e seu resultado é similar ao da regressão linear simples: a estimativa dos parâmetros por mínimos quadrados, r^2, uma razão-F para testar a significância do modelo geral, as variâncias do erro, os intervalos de confiança e os testes de hipóteses para cada um dos coeficientes. A análise dos resíduos e o teste de dados discrepantes e pontos influenciáveis podem ser realizados como na regressão linear simples. Os métodos bayesianos e de Monte Carlo também podem ser usados para modelos de regressão múltipla.

No entanto, um novo problema surge ao avaliar modelos de regressão múltipla, o que não aconteceu na regressão linear simples: pode haver correlação entre as variáveis preditoras, conhecida como **multicolinearidade** (*ver* Graham, 2003). Idealmente, as variáveis preditoras são **ortogonais** umas às outras, uma vez que todos os valores de uma variável preditora são encontrados em combinação com os da segunda. Quando discutimos o delineamento de experimentos com dois fatores no Capítulo 7, enfatizamos a importância de garantir que todas as com-

binações de tratamentos possíveis fossem representadas em um delineamento totalmente cruzado.

O mesmo princípio se aplica à regressão múltipla: idealmente, as variáveis preditoras não devem ser correlacionadas umas com as outras. Quando isso acontece, dificulta a separação das contribuições de cada uma para a variável resposta. Matematicamente, as estimativas por mínimos quadrados também começam a ficar instáveis e difíceis de calcular se há muita multicolinearidade entre as variáveis preditoras. Sempre que possível, delineie seu estudo para evitar correlações entre as variáveis preditoras. Contudo, em um estudo observacional, você pode não ser capaz de quebrar a covariância de suas variáveis preditoras, ou seja, terá que aceitar certo nível de multicolinearidade. Uma cuidadosa análise dos resíduos e diagnósticos é uma estratégia para lidar com a covariância entre as variáveis preditoras. Outra estratégia é combinar matematicamente um conjunto de variáveis preditoras intercorrelacionadas em um número menor de variáveis ortogonais, usando métodos multivariados, como os componentes principais ou a análise discriminante (*ver* Capítulo 12).

A multicolinearidade é um problema no estudo das formigas? Realmente não. Há muito pouca correlação entre altitude e latitude, as duas variáveis preditoras para a riqueza de espécies de formigas (Figura 9.14).

Análise de caminhos

Todos os modelos de regressão que discutimos até agora (linear, robusta, quantílica, não linear e múltipla) começam com a designação de uma única variável resposta e

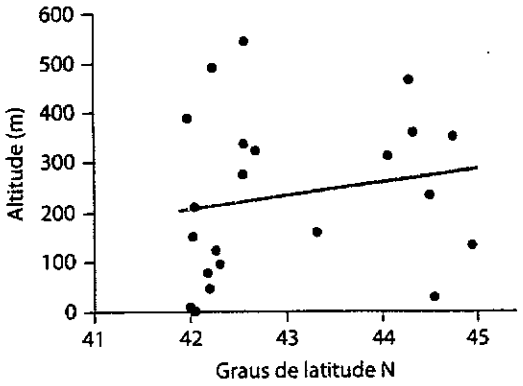


Figura 9.14 A falta de colinearidade entre as variáveis preditoras fortalece as análises de regressão múltipla. Cada ponto representa a latitude e a altitude em um conjunto de 22 parcelas de floresta em que a riqueza de espécies de formigas foi medida (Figura 9.12). Embora a altitude e a latitude possam ser usadas simultaneamente como preditoras em um modelo de regressão múltipla (Figura 9.13), o ajuste do modelo pode ser comprometido se houver fortes correlações entre as variáveis preditoras (colinearidade). Neste caso, há somente uma correlação muito fraca entre as variáveis preditoras latitude e altitude ($r^2 = 0{,}032$; $F_{1,20} = 0{,}66$, $P = 0{,}426$). É sempre uma boa ideia testar as correlações entre as variáveis preditoras ao usar a regressão múltipla. (Detalhes sobre os dados e amostragem em Gotelli & Ellison, 2002a,b.)

Figura 9.15 Análise de caminhos para modelos de colonização passiva e de cadeia alimentar dos inquilinos da *Sarracenia*. Cada balão representa um táxon diferente que pode ser encontrado nas folhas de *Sarracenia purpurea*. Os "insetos-presa" capturados formam as bases dessa cadeia alimentar complexa, que inclui vários níveis tróficos. Os dados subjacentes consistem das abundâncias dos organismos medidos em folhas que possuem diferentes níveis de água em cada uma (Gotelli et al., dados não publicados). Os tratamentos foram aplicados durante uma estação (maio a agosto de 2000) com um experimento do tipo pressão (*ver* Capítulo 6) em 50 plantas ($n = 118$ folhas). Os dados de abundância foram ajustados a modelos diferentes de caminhos, que representam dois modelos diferentes de organização da comunidade. (A) No modelo de colonização passiva, as abundâncias de cada táxon dependem do volume de água em cada folha e da quantidade de presas nas folhas, mas nenhuma interação entre os táxons é evocada. (B) No modelo trófico, as abundâncias são totalmente determinadas pelas interações tróficas entre os inquilinos. O número associado a cada seta é o coeficiente padronizado do caminho. Os coeficientes positivos indicam que o aumento na abundância das presas leva a uma elevação na abundância de predadores. Os coeficientes negativos indicam que o aumento da abundância dos predadores leva ao decréscimo na abundância de presas. A espessura de cada seta é proporcional ao tamanho do coeficiente. Os coeficientes positivos são representados por linhas verdes e os negativos por linhas pretas.

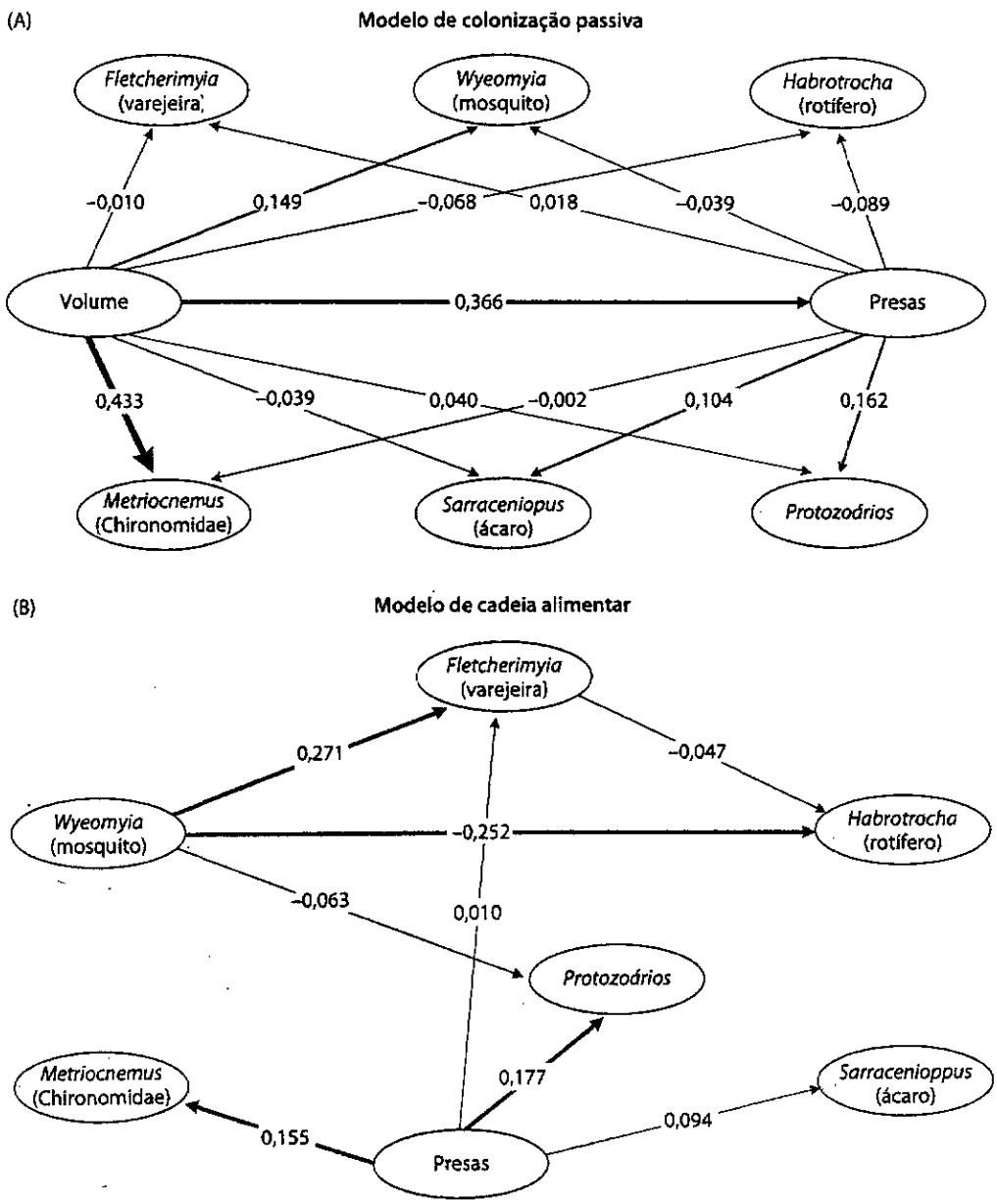
uma ou mais variáveis preditoras que podem explicar a variação na resposta. Mas, na realidade, muitos modelos de processos ecológicos não organizam as variáveis em uma única variável resposta e múltiplas variáveis preditoras. Em vez disso, as variáveis medidas podem agir simultaneamente umas sobre as outras com relações de causa e efeito. Em vez de isolar a variação em uma única variável, deveríamos tentar explicar o padrão geral de covariância de um conjunto de variáveis contínuas.

Este é o objetivo da **análise de caminhos** (*path analysis*). Ela força o usuário a especificar um diagrama de caminhos que ilustra as relações hipotéticas entre as variáveis, que são conectadas umas às outras por setas simples ou duplas. Aquelas que não interagem diretamente não são conectadas por setas. Um diagrama de caminhos deste tipo representa uma hipótese mecanística de interações em um sistema de variáveis.[16] Ele também representa uma hipótese estatística acerca da estrutura da matriz de variância e covariância dessas variáveis (*ver* Nota de rodapé 5, na página 246). Os parâmetros parciais da regressão podem ser estimados para cada caminho individual do diagrama, e um resumo de estatísticas de bondade do ajuste (*goodness of fit*) pode ser derivado para a avaliação geral do modelo.

Por exemplo, a análise de caminhos pode ser usada para testar diferentes modelos de estrutura das comunidades de espécies de invertebrados que

Sewall Wright

[16] A análise de caminhos foi usada primeiramente pelo geneticista de populações Sewall Wright (1889-1988) como um método para analisar os padrões de covariância na herança de caracteres. Ela tem sido usada há bastante tempo nas ciências sociais, mas se tornou popular aos ecólogos apenas recentemente. Ela tem algumas das mesmas potencialidades e fraquezas da análise de regressão múltipla. *Ver* Kingsolver & Schemske (1991), Mitchell (1992), Petraitis et al. (1996) e Shipley (1997) para boas discussões sobre análise de caminhos em ecologia e evolução.



co-ocorrem em folhas da planta-jarro (Gotelli et al., dados não publicados). Os dados consistem de repetidos censos das comunidades inteiras que vivem em folhas dessa planta (n = 118 folhas). As réplicas são folhas individuais em que se fez o censo, e as variáveis contínuas são as abundâncias de cada espécie de invertebrados.

Um modelo para explicar a estrutura da comunidade é o de colonização passiva (Figura 9.15A), no qual a abundância de cada espécie é amplamente determi-

nada pelo volume de água nas folhas e pelos recursos alimentares (formigas) disponíveis nas folhas. Esse modelo não inclui nenhuma interação explícita entre as espécies residentes, embora elas partilhem dos mesmos recursos alimentares. Um segundo modelo para explicar a estrutura da comunidade é o de cadeia alimentar, no qual as interações tróficas entre as espécies são importantes mecanismos reguladores de suas abundâncias (Figura 9.15B). As presas estão na base da cadeia trófica e são consumidas por ácaros, chironomídeos*, protozoários e moscas varejeiras. Embora estas sejam os predadores de topo, algumas das espécies desta comunidade possuem múltiplos papéis tróficos.

Os modelos de colonização passiva e de cadeia alimentar representam duas hipóteses *a priori* acerca das forças que controlam a abundância dos inquilinos. A Figura 9.15 ilustra os **coeficientes de caminhos** associados a cada modelo. No modelo de colonização passiva, os maiores coeficientes são os positivos entre volume da folha com abundância de chironomídeos, volume da folha com abundância de presas e abundância de presas com a de protozoários. Os intervalos de confiança da maioria dos coeficientes englobam o zero.

No modelo de cadeia alimentar, existem altos coeficientes positivos entre presas e chironomídeos, entre presas e protozoários e entre abundância de mosquitos e de varejeiras. Um forte coeficiente negativo liga os mosquitos e suas presas, os rotíferos. Como no modelo de colonização passiva, os intervalos de confiança de vários coeficientes englobam o zero.

A análise de caminhos é praticamente uma abordagem bayesiana de modelar e pode ser implementada em uma estrutura completamente bayesiana, na qual os coeficientes de caminhos são tratados como variáveis aleatórias (Congdon, 2002). Ela requer que o usuário especifique uma hipótese *a priori* na forma de diagrama de caminhos com setas de causa e efeito. Em contraste, as regressões simples e múltiplas têm o espírito mais frequentista: propomos um modelo linear parcimonioso para os dados e testar a simples hipótese nula de que os coeficientes do modelo (β_i) não diferem de zero.

CRITÉRIOS PARA SELEÇÃO DE MODELOS

A análise de caminhos e a regressão múltipla apresentam um problema que ainda não havia surgido em nossa incursão pelos métodos estatísticos: como decidir entre modelos alternativos. Tanto na regressão múltipla quanto na análise de caminhos, os coeficientes e a significância estatística podem ser calculados para qualquer modelo individual. Mas como escolhemos entre os diferentes modelos candidatos? Para a regressão múltipla, pode haver tantas variáveis preditoras quanto pudermos usar. Por exemplo, no estudo das formigas de floresta, também medimos a estrutura da vegetação e a cobertura do dossel, que potencialmente poderiam afetar a riqueza de espécies de formigas. Devemos usar todas as variáveis, ou há alguma forma de selecionar um subconjunto de variáveis

*N. de T. No original *midges*, do gênero *Mitriocnemus*, insetos da família Chironomidae. A tradução literal seria "vermes", porém, no Brasil, não costumamos denominar os chironomídeos assim.

preditoras para explicar com mais parcimônia a variação na variável resposta? Para a análise de caminhos, como decidir qual dos modelos *a priori* melhor se ajusta aos dados?

Métodos de seleção de modelos para regressão múltipla

Para um conjunto de dados que tem uma variável resposta e n variáveis preditoras existem $(2^n - 1)$ modelos de regressão possíveis que podem ser criados. Isto varia desde os modelos lineares simples, com somente uma variável preditora, até o modelo completo, saturado, de regressão múltipla, que tem n variáveis preditoras presentes.[17]

Pode parecer que o melhor modelo seria o com o maior r^2 e que explica mais da variação dos dados. Entretanto, você pode observar que o modelo completo, saturado, sempre tem o com maior r^2. Adicionar variáveis em um modelo de regressão nunca aumenta a *SQR* — geralmente a diminui —, embora o decréscimo possa ser muito pequeno ao adicionar certas variáveis ao modelo.

Por que não usar apenas as variáveis nas quais os coeficientes de inclinação (β_i) do modelo saturado são estatisticamente diferentes de zero? O problema, neste caso, é que os coeficientes – e sua significância estatística – dependem de quais variáveis são incluídas no modelo. Não há nenhuma garantia de que o modelo reduzido contendo apenas os coeficientes de regressão significativos será necessariamente o melhor. De fato, alguns coeficientes no modelo reduzido poderão não ser estatisticamente significativos, em especial se houver multicolinearidade entre as variáveis.

As estratégias para seleção de modelo incluem a **seleção progressiva** (*forward selection*), a **eliminação regressiva** (*backward elimination*) e os **métodos gradativos** (*stepwise methods*). Na seleção progressiva adicionamos variáveis, uma de cada vez, e paramos após certo critério. Na eliminação regressiva de modelos, iniciamos com o modelo saturado completo (todas as variáveis incluídas) e passamos a eliminar uma variável de cada vez. Os modelos gradativos incluem tanto passos de seleção progressiva quanto de eliminação regressiva, para fazer comparações de troca, tirando e colocando variáveis no modelo, e avaliando as mudanças no critério.

Tipicamente, dois critérios são usados para decidir quando parar de adicionar (ou remover) variáveis. O primeiro critério é a mudança na estatística-F do modelo ajustado. A razão-F é uma escolha melhor que o r^2 ou a *SQR*, pois uma

[17] Na verdade, existem bem mais modelos possíveis que estes. Por exemplo, se tivermos duas variáveis preditoras, X_1 e X_2, podemos criar uma variável composta X_1X_2. O modelo de regressão

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

possui coeficientes tanto para os efeitos principais de X_1 e X_2 como para a interação entre X_1 e X_2. Este termo é análogo à interação entre variáveis discretas na ANOVA com dois fatores (*ver* Capítulo 10). Como dissemos, os pacotes estatísticos aceitarão qualquer modelo linear que você desejar construir, sendo responsabilidade sua examinar os resíduos e considerar as implicações biológicas das diferentes estruturas dos modelos.

mudança na estatística-F depende tanto da redução no r^2 quanto do número de parâmetros que é incluído no modelo. Na seleção progressiva, continuamos adicionando variáveis até que o aumento na estatística-F seja menor que um limite estabelecido. Na eliminação regressiva, eliminamos variáveis até que haja uma grande queda na estatística-F. O segundo critério para seleção de modelos é a **tolerância**. As variáveis são removidas do modelo ou não são adicionadas caso elas criem muita multicolinearidade entre o conjunto de preditoras. Tipicamente, o algoritmo dos programas cria uma regressão múltipla na qual a candidata é a variável resposta, e as variáveis que já estão no modelo são as preditoras. A quantidade $(1 - r^2)$ neste contexto é a tolerância. Se a tolerância for muito baixa, a nova variável mostra muita correlação com o conjunto existente de variáveis preditoras e, por isso, não será incluída na equação.

Todos os programas de computador contêm pontos de corte já configurados para os valores de F e de tolerância, motivo pelo qual podem gerar um conjunto razoável de variáveis preditoras para a regressão múltipla. No entanto, existem dois problemas nesta abordagem. O primeiro é que não há nenhuma razão teórica do porque os métodos de seleção de variáveis encontrarão o melhor subconjunto de variáveis preditoras na regressão múltipla. O segundo problema é que esses métodos de seleção fundamentam-se em um critério de otimização, de forma que você termine com apenas um melhor modelo. Os algoritmos não lhe dão modelos alternativos que podem ter valores de F e estatísticas da regressão que sejam muito similares. Esses modelos alternativos incluem diferentes variáveis, mas são estatisticamente indistinguíveis do único modelo de melhor ajuste.

Se o número de variáveis candidatas não for muito grande (< 8), uma boa estratégia é calcular todos os modelos de regressão possíveis e então avaliá-los com base na estatística-F ou em outro critério de mínimos quadrados. Desta forma, você pode no mínimo ver se existe um grande conjunto de modelos similares que podem ser estatisticamente equivalentes. Você também pode encontrar padrões nas variáveis – talvez uma ou duas variáveis sempre apareçam no conjunto final de modelos – que lhe ajudem a escolher o subconjunto final. Você não pode confiar em análises estatísticas automatizadas ou computadorizadas para selecionar o conjunto de variáveis correlacionadas e identificar o modelo correto para os seus dados. Os métodos de seleção de modelos que existem são razoáveis, mas arbitrários.

Métodos de seleção de modelos na análise de caminhos

A análise de caminhos força o investigador a propor modelos específicos que posteriormente são avaliados e comparados. A seleção do "melhor" modelo na análise de caminhos presume que o "correto" é incluído entre as alternativas em avaliação. Se nenhum dos modelos de caminhos propostos for correto, você simplesmente escolherá um incorreto que se ajuste relativamente bem, comparado aos outros modelos incorretos alternativos.

Tanto na análise de caminhos quanto na regressão múltipla, o problema é escolher um modelo com variáveis suficientemente capazes de explicar, de forma

correta, a variação nos dados e minimizar a soma dos quadrados dos resíduos, mas que não tenha tantas variáveis de forma que você ajuste equações ao ruído aleatório dos dados. As **estatísticas de critérios de informação** fazem um balanço entre a redução na soma dos quadrados e a adição de parâmetros ao modelo. Esses métodos penalizam os modelos com mais parâmetros, mesmo se esses modelos inevitavelmente reduzem a soma dos quadrados. Por exemplo, um critério de informação para a regressão múltipla é o r^2 **ajustado**:

$$r^2_{aju} = 1 - \left(\frac{n-1}{n-p}\right)\left(1-r^2\right) \qquad (9.45)$$

onde n é o tamanho amostral e p é o número de parâmetros no modelo (p = 2 na regressão linear simples, com o parâmetro da inclinação e o do intercepto). Como o r^2, esse número aumenta quanto mais a variância residual for explicada. Entretanto, o r^2_{adj} também decresce conforme mais parâmetros são adicionados ao modelo. Portanto, o modelo com o maior r^2_{adj} não será necessariamente o modelo com mais parâmetros. De fato, se os modelos têm muitos parâmetros e os dados são mal ajustados, o r^2_{adj} pode até mesmo se tornar negativo.

Para a análise de caminhos, o **critério de informação de Akaike**, ou **AIC** (*Akaike information criterion*), pode ser calculado assim:

$$\text{AIC} = -2\log\left[L(\hat{\theta}\,|\,dados)\right] + 2K \qquad (9.46)$$

onde $L(\hat{\theta}\,|\,y)$ é a verossimilhança do parâmetro estimado do modelo ($\hat{\theta}$), de acordo com os dados, e K é o número de parâmetros no modelo. Na análise de caminhos, o número de parâmetros no modelo não é simplesmente o de setas no diagrama de caminhos. A medida de dispersão de cada variável também precisa ser estimada.

O AIC pode ser visto como uma medida de "ruindade do ajuste", pois, quanto maior o número, pior os dados se ajustam à estrutura da matriz de variância e covariância implícita no diagrama de caminhos. Por exemplo, o modelo de colonização aleatória dos inquilinos possui 21 parâmetros e um índice de validação cruzada (uma medida do AIC) de 0,702, enquanto o modelo de cadeia alimentar possui 15 parâmetros e um índice de validação cruzada de 0,466.

Seleção de modelos bayesianos

As análises bayesianas podem ser usadas para comparar diferentes hipóteses e modelos (como os de regressão com ou sem o termo de inclinação β_1; *ver* Equações 9.20 e 9.21). Dois métodos foram desenvolvidos (Kass & Raftery, 1995, Congdon, 2002). O primeiro, chamado de **fatores de Bayes**, estima a verossimilhança referente a uma hipótese ou um modelo, relativa à outra hipótese ou modelo. Como em uma análise bayesiana completa, temos probabilidades *a priori* $P(H_0)$ e $P(H_1)$ para nossas hipóteses ou modelos. A partir do Teorema de Bayes (*ver* Capítulo 1),

as probabilidades *a posteriori* $P(H_0|\text{dados})$ e $P(H_1|\text{dados})$ são proporcionais às probabilidades *a priori* × as verossimilhanças $L(\text{dados}|H_0)$ e $L(\text{dados}|H_1)$, respectivamente:

$$P(H_i|\text{dados}) \propto L(\text{dados}|H_i) \times P(H_i) \quad (9.47)$$

Definimos a **razão de chances *a priori*** (*prior odds ratio*), ou a probabilidade relativa de uma hipótese *versus* outra hipótese, como:

$$P(H_0)/P(H_1) \quad (9.48)$$

Se estas forem as duas únicas alternativas, então $P(H_0) + P(H_1) = 1$ (pelo Primeiro Axioma da Probabilidade, Capítulo 1). A Equação 9.48 expressa quão mais ou menos provável uma hipótese é em relação à outra antes de conduzir um experimento. No início do estudo, esperamos que as probabilidades *a priori* de cada hipótese ou modelo sejam grosseiramente iguais (caso contrário, por que gastar seu tempo obtendo dados?), assim a Equação 9.48 ≈ 1.

Após a coleta dos dados, calculamos as probabilidades *a posteriori*. O fator de Bayes é a **razão de chances *a posteriori***:

$$P(H_0|\text{dados}) / P(H_1|\text{dados}) \quad (9.49)$$

Se a Equação 9.49 $\gg 1$, podemos ter razão para acreditar que a H_0 é favorecida em relação à H_1; contudo, se a Equação 9.49 $\ll 1$, podemos confirmar a H_1. Usamos as Equações 9.47 e 9.48 para, em conjunto, obter a razão de chances *a posteriori*:

$$P(H_0|\text{dados}) / P(H_1|\text{dados}) = [L(\text{dados}|H_0) / L(\text{data}|H_1)] \times [P(H_0) / P(H_1)] \quad (9.50)$$

A Equação 9.50 diz que a razão de chances *a posteriori* é igual à de verossimilhanças vezes a razão de chances *a priori*. Como esta é igual a 1 em um experimento com priores não informativas [$P(H_0)/ P(H_1) = 0{,}5$], a razão de verossimilhanças (também chamada de fator de Bayes) pode ser usada como uma estimativa da razão de chances *a posteriori*.[18]

O segundo método bayesiano para comparar modelos é aproximar os fatores de Bayes quando as priores não são informativas. Esse método, chamado de **critério de informação de Bayes (BIC – *Bayesian information criterion*)** é usado para seleção de modelos, como escolher entre modelos de regressão. Permita que λ = razão de verossimilhança $L(\text{dados} \mid M_0) / L(\text{dados} \mid M_1)$ para dois modelos

[18] Os fatores de Bayes são mais úteis quando existem priores informativas. Como, agora, a maioria dos programas bayesianos usa priores não informativas, há muito menos ênfase nos fatores de Bayes do que quando o uso de priores informativas era a norma. *Ver* Kass & Raftery (1995) para uma discussão adicional sobre os fatores de Bayes.

M_0 e M_1, cada um com diferentes números de parâmetros p_1 e p_2 (por exemplo, a Equação 9.20 tem um parâmetro, β_0, e a Equação 9.21 tem dois parâmetros, β_0 e β_1). Definimos o BIC como sendo:

$$\text{BIC} = 2\log_e \lambda - (p_1 - p_2)\log_e n \qquad (9.51)$$

onde n é o tamanho amostral. O melhor modelo é o que tiver o menor valor de BIC.[19]

Em resumo, a seleção de um conjunto reduzido de variáveis é uma atividade comum em estudos de regressão múltipla, mas os critérios usados são um tanto arbitrários. Tais análises sofrem quando há multicolinearidade nos dados e se o tamanho amostral é muito pequeno. Em geral, existirão no mínimo de 10 a 20 observações para cada variável independente que você está considerando. Muitos estudos de regressão, na realidade, não possuem replicação adequada para a regressão múltipla, e há o perigo de que, mesmo com métodos gradativos, os modelos resultantes não sejam nem ótimos nem parcimoniosos.

A regressão e a análise de caminhos com estatística de AIC ou de BIC estão um passo adiante a esse respeito, pois elas forçam o investigador a considerar estruturas de modelos explícitas. Entretanto, esse tipo de estrutura de trabalho nem sempre pode estar disponível, em especial nos estágios iniciais de um estudo. Burnham & Anderson (2002) constituem uma excelente introdução ao tópico geral de ajuste de modelos e critérios de informação.

RESUMO

A regressão é uma ferramenta estatística poderosa para avaliar a relação entre duas ou mais variáveis contínuas. Soluções de parâmetros por mínimos quadrados são apropriadas e sem viés se o modelo realmente for linear, se a variável X for medida sem erros, se as observações forem independentes e se os termos do erro tiverem distribuição normal. A análise de resíduos é um passo-chave para avaliar a validade das premissas da regressão. Métodos mais avançados incluem a regressão robusta para tratar dados discrepantes, regressão logística e não linear para relações funcionais que não são lineares, e regressão múltipla e análise de caminhos para tratar múltiplas variáveis preditoras e estruturas complexas de modelos *a priori*.

[19] O uso do BIC é similar ao uso do critério de informação de Akaike (AIC) na seleção de modelos paramétricos (Burnham & Anderson 2002), mas as propriedades estatísticas do BIC não são tão bem comportadas como as do AIC.

10 Análise de Variância

A análise de variância, ou **ANOVA**, é a técnica de Fisher para partição da soma dos quadrados, algo que vimos em nossa discussão sobre regressão (Capítulo 9). Em geral, a ANOVA se refere a uma classe de delineamentos amostrais ou experimentais, na qual a variável preditora é categórica e a variável resposta é contínua. Exemplos desses delineamentos incluem os de um fator, os blocos aleatorizados e os de parcelas subdivididas. No Capítulo 7, descrevemos o formato físico desses delineamentos, a razão para utilizá-los e algumas vantagens e desvantagens de cada um. Este capítulo se concentra na análise de dados e testes de hipóteses que são associados a cada um desses delineamentos.

Primeiro, explicamos a mecânica da partição da soma dos quadrados – uma técnica fundamental da ANOVA. A seguir, traçamos as premissas da ANOVA. Se cumpridas, podemos usar a razão-F de Fisher para estimar valores de *P* para a soma dos quadrados partidas. Para cada um dos delineamentos do Capítulo 7, explicamos os modelos subjacentes e as tabelas de ANOVA associadas. Também descrevemos como plotar os dados de forma que você possa entender os termos de interação em ANOVAs de dois fatores (*two-way*) e ANCOVAs. Explicamos como incorporar fatores aleatórios e fixos nas suas análises e como usar esses modelos para fazer a partição de variâncias de um conjunto de dados. Descrevemos alguns procedimentos usados após seu teste de ANOVA básico para comparar médias (contrastes *a priori* e comparações *a posteriori*) e concluímos com uma discussão sobre como interpretar um conjunto de valores *P* coletados em experimentos múltiplos.

É fácil perder o foco dos objetivos da ANOVA: a comparação de médias entre grupos que foram amostrados aleatoriamente. O Capítulo 5 detalha um exemplo no qual um pesquisador deseja comparar a densidade de formigueiros em hábitats de floresta e campo. Neste sentido, a ANOVA é simplesmente uma extensão do conhecido teste-*t*, que é usado para comparar as médias de dois grupos amos-

trados. Se você ainda não o tiver feito, leia os Capítulos 7 e 9 antes deste. Embora o Capítulo 9 discuta regressão, ele também apresenta a ideia de modelo linear, efeitos de tratamentos, partição da soma de quadrados e do arranjo de uma tabela de ANOVA. De fato, tanto a regressão quanto a ANOVA são casos especiais de um modelo linear generalizado (McCullagh & Nelder, 1989). Embora os programas estatísticos solucionem todas as equações apresentadas neste capítulo, é importante que você entenda este conteúdo, pois as configurações-padrão usadas por muitos programas não irão gerar as análises corretas para muitos delineamentos experimentais comuns.

SÍMBOLOS E RÓTULOS EM ANOVA

Uma grande dor de cabeça ao interpretar tabelas de ANOVA é entender os símbolos e rótulos convencionais. Existem muitas variáveis para analisarmos e não há um conjunto de notações consistentes na literatura.

Utilizamos um sistema relativamente simples. Primeiro, o símbolo Y é sempre reservado para a variável resposta medida, da mesma forma que vem sendo usado ao longo do livro. O símbolo $\overline{Y}$ indica a média geral dos dados. Qualquer média que seja calculada para um subgrupo particular é indicada com um subscrito, como $\overline{Y}_i$. Uma variável subscrita que não possua uma barra acima indica um dado em particular, como Y_{ij}. A variável μ indica o valor esperado de uma variável em um modelo, e o termo residual do erro é indicado por ε, geralmente com algum subscrito para indicar os componentes de tratamentos diferentes.

Letras maiúsculas *A, B, C...* designam os diferentes fatores de um modelo. Os níveis diferentes das variáveis são indicados por subscritos *i, j, k...* Por exemplo, poderíamos escrever A_i para indicar o nível *i* do Fator *A*, e B_j para indicar o nível *j* do Fator B. O número máximo de níveis para um fator é indicado pela *letra minúscula* correspondente. Então, se A_i indica o nível *i* do tratamento para o Fator *A*, o número de níveis do Fator *A* varia de $i = 1$ até a. A exceção a esse padrão é o *n* minúsculo, que representa o número de réplicas utilizadas para estimar a soma dos quadrados dentro de grupos (a soma dos quadrados dos resíduos), o nível mais baixo em que as réplicas são realizadas. Sempre utilizaremos o símbolo padrão da variância σ^2, independente de quando o componente da variância é para um fator fixo ou para um aleatório. Mostramos com clareza, no texto ou na legenda de tabela, quando cada fator no modelo é fixo ou aleatório.

ANOVA E PARTIÇÃO DA SOMA DOS QUADRADOS

A análise de variância é construída no conceito de partição da soma dos quadrados, que foi introduzido no Capítulo 9. Rapidamente, a variação total em um conjunto de dados pode ser expressa como uma soma de quadrados: a diferença entre cada observação (Y_i) e a média geral dos dados ($\overline{Y}$) é elevada ao quadrado e somada. Essa variação total pode ser repartida ou dividida em diferentes componentes.

Alguns componentes representam variação aleatória ou variação de erro que não é atribuível a nenhuma causa específica; ela pode resultar do erro de observação ou outras forças não específicas (*ver* Notas de rodapé 10 e 11, Capítulo 1). Outros componentes representam os efeitos dos tratamentos experimentais aplicados às réplicas, ou as diferenças entre categorias amostrais. Uma análise estatística envolve especificar um modelo subjacente de como as observações podem ser afetadas por diferentes tratamentos, fazer a partição da soma dos quadrados entre os diferentes componentes do modelo e utilizar, então, os resultados para testar hipóteses estatísticas sobre força dos efeitos em particular.

Ilustramos a partição da soma dos quadrados com um delineamento de ANOVA de um fator (*one-way*) utilizado para testar os efeitos do derretimento precoce da neve sobre o crescimento de plantas alpinas (p. ex., Price & Waser, 1998; Dunne et al., 2003). Tal experimento pode ter 3 grupos de tratamento e 4 réplicas por tratamento ($4 \times 3 = 12$ observações no total). Quatro parcelas são sem manipulação, não tendo sido feitas mudanças nas parcelas além daquelas que ocorrem durante o censo. Quatro parcelas são aquecidas com bobinas de aquecimento permanente a energia solar que derretem a neve da primavera mais cedo que o normal. Quatro parcelas adicionais atuam como controles: elas são preenchidas com bobinas de aquecimento que nunca são ativadas. Após 3 anos de aplicação do tratamento, você mede o comprimento do período de floração da esporinha (*Delphinium nuttallianum*) em cada parcela.

Os resultados são mostrados na Tabela 10.1, junto com todos os cálculos necessários para esse exemplo. Embora a maioria dos cálculos da ANOVA hoje seja feita em um computador, é importante trabalhar à mão nesse exemplo, para você ter um entendimento sólido de como a soma dos quadrados é dividida.

Como descrevemos nos Capítulos 3 e 9, começamos nossas análises calculando a soma dos quadrados total dos dados, que é a soma dos desvios quadrados de cada observação (Y_i) para média geral ($\bar{Y}$). Com um fator, há de $i = 1$ até a tratamentos e de $j = 1$ à n réplicas por tratamento, com um total de $a \times n$ observações. Em nosso exemplo, há $a = 3$ tratamentos (sem manipulação, controle e tratamento) e $n = 4$ réplicas por tratamento, com um tamanho amostral de $a \times n = 3 \times 4 = 12$. Então, podemos escrever:

$$SQ_{total} = \sum_{i=1}^{a} \sum_{j=1}^{n} (Y_{ij} - \bar{Y})^2 \qquad (10.1)$$

A soma dos quadrados total para os dados da Tabela 10.1 é 41,66.

Essa soma dos quadrados total reflete o desvio de cada observação da média geral. Ela pode ser decomposta (dividida) em duas fontes diferentes. O primeiro **componente de variação** é a **variação entre grupos**. A variação entre grupos representa diferenças entre as médias de cada grupo de tratamento. Pensando na média de cada grupo com uma única observação, essa fonte de variação é:

TABELA 10.1 Fazendo a partição da soma dos quadrados na ANOVA

Sem manipulação	Controle	Tratamento
10	9	12
12	11	13
12	11	15
13	12	16
$\bar{Y}_1 = 11{,}75$	$\bar{Y}_2 = 10{,}75$	$\bar{Y}_3 = 14{,}00$
$\sum_{j=1}^{n}(Y_{1j}-\bar{Y}_1)^2 = 4{,}75$	$\sum_{j=1}^{n}(Y_{2j}-\bar{Y}_2)^2 = 4{,}75$	$\sum_{j=1}^{n}(Y_{3j}-\bar{Y}_3)^2 = 10{,}00$
$\sum_{j=1}^{n}(\bar{Y}_1-\bar{Y})^2 = 0{,}68$	$\sum_{j=1}^{n}(\bar{Y}_2-\bar{Y})^2 = 8{,}08$	$\sum_{j=1}^{n}(\bar{Y}_3-\bar{Y})^2 = 13{,}40$
$\sum_{j=1}^{n}(Y_{1j}-\bar{Y})^2 = 5{,}43$	$\sum_{j=1}^{n}(Y_{2j}-\bar{Y})^2 = 12{,}83$	$\sum_{j=1}^{n}(Y_{3j}-\bar{Y})^2 = 23{,}40$

$$SQ_{entre\ grupos} = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{Y}_i - \bar{Y})^2 \qquad (10.2)$$

Esta equação contém dois somatórios, um para os grupos de tratamento a e um para as n_i observações em cada grupo de tratamento. Fazer o primeiro somatório é fácil: pegue a média de cada tratamento, subtraia da média geral, eleve ao quadrado e some os termos para cada um dos grupos de tratamento a. Mas como "somaremos os j" quando não há um subscrito j na equação $(\bar{Y}_i - \bar{Y})^2$? Como há n_i observações em cada grupo de tratamento, você apenas multiplica o primeiro somatório pela constante n_i. Então, a Equação 10.2 é equivalente a:

$$SQ_{entre\ grupos} = \sum_{i=1}^{a} n_i(\bar{Y}_i - \bar{Y})^2$$

Nos delineamentos balanceados que discutimos neste capítulo, n_i é o mesmo para todos os grupos de tratamentos ($n_i = n$), e isto simplifica para:

$$SQ_{entre\ grupos} = n\sum_{i=1}^{a}(\bar{Y}_i - \bar{Y})^2$$

$$\bar{Y} = 12{,}17$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2 = 19{,}50 = SQ_{dentro\ de\ grupos}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{Y}_i - \bar{Y})^2 = 22{,}16 = SQ_{entre\ grupos}$$

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y})^2 = 41{,}66 = SQ_{total}$$

Os dados brutos hipotéticos consistem em observações do período de floração (em semanas) da esporinha em um conjunto de 12 parcelas em prados alpinos. Quatro dessas parcelas receberam um tratamento experimental de aquecimento, quatro parcelas serviram como controles (os aparelhos de aquecimento montados, mas não são ligados) e quadro parcelas não foram manipuladas. Existem de $i = 1$ a 3 grupos (= tratamentos), com $n = 4$ réplicas por grupos. Esta tabela ilustra os cálculos básicos da ANOVA e a partição da soma dos quadrados. Na primeira linha dos cálculos, temos a média geral $\bar{Y}$ para todas as 12 réplicas (12,17) e a média $\bar{Y}_i$ para cada grupo de tratamento (11,75, 10,75 e 14,00). Na segunda linha de cálculos, somamos o desvio quadrado de cada observação da média do próprio grupo $(Y_{ij} - \bar{Y}_i)^2$ (4,75, 4,75 e 10,00) e somamos estes termos para obter a soma dos quadrados dentro de grupos (19,50). Na terceira linha de cálculos, determinamos o desvio quadrado da média de cada grupo para a média geral $(\bar{Y}_i - \bar{Y})^2$ multiplicada pelo tamanho amostral $n = 4$ (0,68, 8,08 e 13,40) e somamos estes termos para obter a soma dos quadrados entre grupos (22,16). Na quarta linha de cálculos, determinamos o desvio quadrado de cada observação para a média geral (5,43, 12,83 e 23,40) e somamos estes termos para obter a soma dos quadrados total. Assim, a soma dos quadrados total (41,66) pode ser aditivamente particionada no componente dentro de grupo (19,50) e entre grupos (22,16). Trata-se de uma propriedade algébrica fundamental, útil para qualquer conjunto de números. Usar tais dados para conduzir um teste estatístico faz sentido apenas se a amostragem cumprir os pressupostos gerais da ANOVA e se adequar a um delineamento particular específico de um modelo de ANOVA.

No exemplo apresentado na Tabela 10.1, as três somas dos quadrados para os grupos de tratamentos são 0,17, 2,02 e 3,35. Como há quatro réplicas por tratamento, a soma dos quadrados entre grupos é igual a $4 \times (0{,}17 + 2{,}02 + 3{,}35) = 22{,}16$. No modelo de ANOVA de um fator, os fatores controlados representam processos que acreditamos causar diferenças entre os grupos de tratamentos. O efeito desses fatores são representados pela soma dos quadrados entre grupos (Equação 10.2).

O componente remanescente é a **variação dentro de grupos**. Em vez de calcular o desvio de cada observação para a média geral, calculamos o desvio de cada observação para a média de seu próprio grupo e depois somamos entre os grupos e as réplicas:

$$SQ_{dentro\ de\ grupos} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2 \qquad (10.3)$$

Este componente de variação no conjunto de dados do exemplo é 19,50. A soma dos quadrados dentro de grupos geralmente é chamada de **soma dos quadrados dos resíduos**, de **variação residual** ou de **variação do erro**. Como

na regressão, nos referimos a ela como "resíduo", pois é uma variação que não é explicada pelos fatores controlados, ou experimentais, em nosso modelo. A variação dentro de grupos (Equação 10.3) é descrita como "variação do erro" porque nosso modelo estatístico incorpora esse componente como amostragem aleatória a partir de uma distribuição normal. Em modelos de ANOVA mais complexos, iremos fazer a partição da soma dos quadrados total em componentes múltiplos de variação, cada um representando uma contribuição de um fator no modelo. Em todos os casos, contudo, o que sobra é sempre a soma dos quadrados dos resíduos.

Uma das contribuições-chave de Fisher foi mostrar que os componentes de variação são aditivos:

$$SQ_{total} = SQ_{entre\ grupos} + SQ_{dentro\ de\ grupos} \tag{10.4}$$

Em palavras, a soma dos quadrados total é igual à dos quadrados entre grupos mais a dos quadrados dentro de grupos. Para os dados na Tabela 10.1:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\overline{Y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{Y}_i-\overline{Y})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_i)^2 \tag{10.5}$$

$$41{,}66 \quad = \quad 22{,}16 \quad + \quad 19{,}50$$

Enfatizamos que a partição da soma dos quadrados é uma propriedade puramente algébrica: este resultado servirá para qualquer conjunto de números, independente do que eles representem e de como foram coletados.[1]

Todavia, a partição da soma dos quadrados parece ser uma medida natural dos efeitos dos tratamentos. Se a soma dos quadrados entre grupos é relativamente grande se comparada à dos quadrados dentro de grupos, então as diferenças entre tratamentos podem parecer importantes. Por outro lado, se a soma dos

[1] A prova da partição da soma dos quadrados pode ser rapidamente esboçada. Primeiro, comece com a soma dos quadrados total:

$$SQ_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\overline{Y})^2$$

Agora, vamos adicionar e subtrair $\overline{Y}_i$, de forma que o total não muda:

$$SQ_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}+\overline{Y}_i-\overline{Y}_i)^2$$

Reagrupando os elementos, obtemos os dois componentes familiares da soma dos quadrados:

$$SQ_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left[(Y_{ij}-\overline{Y}_i)+(\overline{Y}_i-\overline{Y})\right]^2$$

Recordando a expansão binomial (ver Capítulo 2), $(a+b)^2 = a^2 + 2ab + b^2$:

(Continua)

quadrados dentro de grupos é grande em relação à dos quadrados entre grupos, concluímos que diferenças entre grupos são fracas ou inconsistentes. Veremos, a seguir, como quantificar essas ideias na análise de variâncias.

OS PRESSUPOSTOS DA ANOVA

Antes de utilizarmos a soma dos quadrados em modelos estatísticos, eles precisam satisfazer o seguinte conjunto de pressupostos.

1. *As amostras são independentes e identicamente distribuídas.* Como sempre, este pressuposto forma a base para qualquer modelo estatístico de amostragem. Assumimos que os dados representam uma amostra aleatória do espaço amostral que você definiu e que as observações dentro e entre tratamentos são independentes entre si (*ver* Capítulo 6). Para todas as descrições das tabelas de ANOVA deste capítulo, assumimos o caso mais simples, no qual o tamanho amostral (n) é o mesmo dentro de todos os grupos. *Ver* Sokal & Rohlf (1995) para uma visão geral de métodos de ANOVA quando os tamanhos amostrais dos grupos de tratamentos são desiguais.
2. *As variâncias são homogêneas entre grupos.* Embora as médias dos grupos amostrados possam diferir entre si, assumimos que a variância dentro de cada grupo é aproximadamente igual àquela dentro de todos os outros grupos. Desta forma, cada grupo de tratamento contribui com aproximadamente o mesmo tanto para a soma dos quadrados dentro de grupos. Na regressão linear, fazemos um pressuposto análogo, de que a variância é homogênea para os diferentes níveis da variável X (*ver* Capítulo 9). Além disso, como na regressão linear, as transformações dos dados com frequência tornaram homogêneas as variâncias (*ver* Capítulo 8).
3. *Os resíduos têm distribuição normal.* Assume-se que os resíduos seguem a distribuição normal, com média igual a zero. Graças ao Teorema do Limite

[1] (*Continuação*)

$$SQ_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_i)^2 + 2\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_i)(\overline{Y}_i - \overline{Y}) + \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{Y}_i - \overline{Y})^2$$

Satisfatoriamente, o segundo termo dessa expansão sempre será igual a zero (pois a soma dos desvios da média = 0; *ver* Capítulo 9) e ficamos com:

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{Y}_i - \overline{Y})^2 + \sum_{i=1}^{a}\sum_{i=1}^{n}(Y_{ij} - \overline{Y}_i)^2$$

A Equação 10.4 pode ser pensada como o Teorema de Pitágoras (*ver* Nota de rodapé 5, Capítulo 12) para distâncias dos dados até as médias e entre as médias. A Equação 10.4 garante que a $SQ_{entre\ grupos}$ e a $SQ_{dentro\ de\ grupos}$ sejam ortogonais entre si. Eles não "fazem sombra" um ro outro e são estatisticamente independentes. Essa independência é necessária para uma interpretação válida das razões-F e ainda é outra vantagem de usar os desvios quadrados para expressar distância.

Central (*ver* Capítulo 2), esse pressuposto não é muito restritivo, especialmente se os tamanhos amostrais são grandes e aproximadamente iguais entre tratamentos, ou se os próprios dados forem médias. Novamente, dados adequadamente transformados em geral podem ter termos do erro com distribuição normal.

4. *As amostras são classificadas corretamente.* Para estudos experimentais, assumimos que todos os indivíduos atribuídos a um tratamento em particular foram tratados de maneira idêntica (p. ex., todos os passarinhos de um tratamento livre de parasitas recebem doses idênticas de antibióticos). Para estudos observacionais, ou experimentos naturais, assumimos que todos os indivíduos agrupados em uma classe particular realmente pertençam àquela classe (p. ex., em um estudo de impacto ambiental, todas as parcelas do estudo foram atribuídas corretamente aos grupos impactados e aos grupos controles). A violação desse pressuposto é potencialmente séria e pode comprometer as estimativas dos valores de *P*. Em estudos de regressão, a premissa análoga é a ausência de erros de medidas na variável *X* (*ver* Nota de rodapé 10, Capítulo 9). Entretanto, os erros de medidas em um delineamento de regressão podem não ser um problema tão sério quanto os de classificação em um delineamento de ANOVA. Um estudo de grande qualidade conduzido com cuidado é a melhor garantia contra erros de classificação e de medida.
5. *Os efeitos principais são aditivos.* Em certos delineamentos de ANOVA, como o de blocos aleatorizados ou de parcelas subdivididas, nem todos os fatores dos tratamentos são completamente replicados. Nesses casos, é necessário assumir que os efeitos principais são estritamente aditivos e que não há interação entre os diferentes fatores. Trataremos desse pressuposto em detalhes mais adiante neste capítulo, quando discutirmos esses delineamentos. As transformações dos dados também podem ajudar a garantir aditividade, em especial quando fatores multiplicativos são transformados logaritmicamente (*ver* Capítulo 3).

TESTES DE HIPÓTESES COM ANOVA

Se os pressupostos da ANOVA são cumpridos (ou não são violados de maneira grave), podemos testar hipóteses com base em um modelo subjacente que é ajustado aos dados. Para a ANOVA de um fator, o modelo é:

$$Y_{ij} = \mu + A_i + \varepsilon_{ij} \qquad \textbf{(10.6)}$$

Neste modelo, Y_{ij} é a réplica j associada ao nível de tratamento i, μ é a média geral verdadeira, ou a média aritmética [$(\overline{Y})$ é a estimativa de μ] e ε_{ij} é o termo do erro. Embora cada observação Y_{ij} possua seu próprio erro ε_{ij} associado, lem-

bre-se de que todos os ε_{ij} são retirados de uma distribuição normal única com uma média 0. O elemento mais importante do modelo é o termo A_i. Este termo representa o componente linear aditivo associado ao nível i do tratamento A. Há um diferente coeficiente A_i associado a cada um dos i níveis de tratamento. Se A_i é um número positivo, o nível de tratamento i tem uma expectativa que é superior à média geral. Se A_i é negativo, a expectativa está abaixo da média geral. Como os A_i representam desvios da média geral e, por definição, a sua soma é igual a zero. A ANOVA nos permite estimar os efeitos do A_i (a média dos tratamentos menos a média geral constituem um estimador imparcial do A_i) e testar hipóteses sobre os A_i.

Qual é a hipótese nula? Se não há nenhum efeito dos tratamentos, então $A_i = 0$ para todos os níveis do tratamento. Portanto, a hipótese nula é:

$$Y_{ij} = \mu + \varepsilon_{ij} \qquad (10.7)$$

Se a hipótese nula é verdadeira, qualquer variação que ocorra entre os grupos de tratamentos (e sempre haverá alguma) reflete o erro aleatório e nada mais.

A tabela de ANOVA fornece um teste geral da hipótese nula de ausência de efeitos do tratamento. A Tabela 10.2 mostra os componentes básicos da tabela de ANOVA para o delineamento de um fator (*ver* "A Anatomia de uma Tabela de ANOVA" no Capítulo 9 para os detalhes e abreviações de uma tabela de ANOVA típica). Começamos calculando o quadrado médio, que é simplesmente a soma dos quadrados dividida pelos seus graus de liberdade correspondentes (*ver* Capítulo 9). Dois quadrados médios são calculados em uma ANOVA de um fator.

O primeiro quadrado médio, para a variação entre grupos, possui $(a - 1)$ graus de liberdade, onde a é o número de tratamentos. O segundo quadrado médio, para a variação dentro de grupos, possui $a(n - 1)$ graus de liberdade. Isso tem sentido intuitivo, porque, dentro de cada grupo, deveria ter $(n - 1)$ graus de liberdade. Com a grupos, temos $a(n - 1)$ graus de liberdade para o quadrado médio dentro de grupos. Além disso, note que os graus de liberdade total = $(a - 1) + a(n - 1) = (an - 1)$, justamente um a menos que o tamanho amostral total. Por que os graus de liberdade não somam o mesmo que o tamanho amostral total (an)? Porque um grau de liberdade é usado para estimar a média geral (μ).

Enquanto o quadrado médio dentro de grupos estima a variância do erro σ^2, o quadrado médio entre grupos estima $\sigma^2 + \sigma^2_A$: a variância do erro *mais* a variância devido aos grupos de tratamento. Portanto, a razão entre essas duas quantidades (QM_{entre}/QM_{dentro}) é um teste apropriado do efeito do tratamento. Quanto maior o efeito do tratamento em relação ao erro quadrado dentro do grupo, maior será a razão-F. Se não há efeito dos tratamentos ($\sigma^2_A = 0$) (a hipótese nula), então o quadrado médio entre grupos = $\sigma^2 + 0 = \sigma^2$, que é o mesmo que o

TABELA 10.2 Tabela de ANOVA para um esquema de um fator

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)	Quadrado médio esperado	Razão-F	Valor de *P*
Entre grupos	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{Y}_i-\overline{Y})^2$	$\frac{SQ_{entre\ grupos}}{(a-1)}$	$\sigma^2+n\sigma_A^2$	$\frac{QM_{entre\ grupos}}{QM_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)$ e $a(n-1)$ graus de liberdade
Dentro de grupos (resíduo)	$a(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_i)^2$	$\frac{SQ_{dentro\ de\ grupos}}{a(n-1)}$	σ^2		
Total	$an-1$	$\sum_{i=1}^{a}\sum_{i=1}^{n}(Y_{ij}-\overline{Y})^2$	$\frac{SQ_{total}}{(an-1)}$	σ_Y^2		

Há a grupos, com n réplicas por grupo. A_i é o efeito do tratamento para o grupo i. O fator de agrupamento pode ser ou fixo ou aleatório. Neste delineamento simples, há apenas uma única razão-F que pode ser construída, a qual testa a hipótese nula de que a variação entre grupos amostrados não é significativamente diferente de 0. A razão-F utiliza o quadrado médio entre grupos no numerador e o quadrado médio dentro de grupos (resíduo) no denominador. *Ver* a Tabela 9.1 para explicações adicionais sobre os elementos de uma tabela de ANOVA.

quadrado médio dentro de grupos, ou a variância do erro σ^2. Neste caso, $QM_{entre} = QM_{dentro}$, e a sua razão, a razão-F, = 1,0. A probabilidade de cauda depende tanto do tamanho da razão-F quanto dos graus de liberdade. Para a ANOVA de um fator, há $(a-1)$ graus de liberdade no numerador e $a\ (n-1)$ graus de liberdade no denominador.

Para os dados da Tabela 10.1, a razão-F é 5,11, com um valor de *P* correspondente de 0,033 (Tabela 10.3). O valor de *P* é pequeno (menor que 0,05), então podemos rejeitar a hipótese nula de ausência de efeito do tratamento. Quando olhamos para os dados brutos (Tabela 10.1), parece ser apropriado rejeitar a hipótese nula: os períodos de floração foram maiores no grupo de tratamento que no grupo controle ou no sem manipulação.

CONSTRUINDO RAZÕES-F

Aqui estão os passos gerais para construir uma razão-F e usá-la para testar hipóteses utilizando ANOVA:

1. Use os quadrados médios associados a um modelo de ANOVA que se ajuste ao seu delineamento amostral ou experimental. Apresentamos os delineamentos básicos no Capítulo 7 e neste (10) daremos as tabelas de ANOVA que os acompanham. Esperamos que siga nosso conselho nos Capítulos 4 e 6 e construa seu modelo *antes* de coletar seus dados!

TABELA 10.3 Tabela de ANOVA de um fator para os dados hipotéticos da Tabela 10.1

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)	Razão-F	Valor de *P*
Entre grupos	2	22,17	11,08	5,11	0,033
Dentro de grupos (resíduo)	9	19,50	2,17		
Total	11	41,67			

As fórmulas na Tabela 10.2 são usadas para calcular as somas dos quadrados, os quadrados médios e a razão-F. A primeira coluna indica a fonte de variação. A segunda coluna indica os graus de liberdade, que são determinados pelo tamanho amostral e pelo número de grupos sob comparação. A terceira coluna fornece as somas dos quadrados, que foram calculadas na Tabela 10.1. A quarta coluna é o quadrado médio, calculado dividindo cada soma dos quadrados pelo seu número de graus de liberdade correspondente. A coluna seguinte fornece a razão-F, calculada como a razão entre os valores dos quadrados médios apropriados. Para uma simples ANOVA de um fator, a única razão-F que pode ser construída testa a variação entre grupos, e a é calculada como (o quadrado médio entre grupos)/(pelo quadrado médio dentro de grupos). O valor de *P* é determinado a partir de um conjunto de tabelas estatísticas para uma razão-F com 2 e 9 graus de liberdade. O valor de *P* indica que há somente 3,3% de chance em obter esses dados (ou outros dados ainda mais extremos) se a hipótese nula for verdadeira. Como esse valor de *P* é menor que o nível padrão do alfa de 0,05, rejeitamos a hipótese nula e concluímos que a variação entre grupos provavelmente é maior do que pode ser explicado somente por chance.

2. Encontre o quadrado médio esperado que inclui o efeito em particular que você está tentando medir e use-o no numerador da razão-F.
3. Encontre o segundo quadrado médio esperado que inclui todos os termos estatísticos no numerador, *com exceção do único termo que você está tentando estimar*, e use-o no denominador da razão-F.
4. Divida o numerador pelo denominador e obtenha sua razão-F.
5. Use tabelas estatísticas ou o resultado do seu computador, determine o valor de probabilidade (valor de *P*) associado à razão-F e seus graus de liberdade correspondentes. A hipótese nula sempre é de que o efeito de interesse é zero. Se a hipótese nula é verdadeira, a razão-F construída adequadamente em geral terá um valor esperado de 1,0. Em contrapartida, se o efeito é muito grande, o numerador será muito maior que o denominador e produzirá uma razão-F que é consideravelmente superior a 1,0.[2]
6. Repita os passos de 2 a 5 para os outros fatores que estiver testando. Uma ANOVA simples de um fator gera apenas uma única razão-F, mas modelos mais complexos permitem que você teste múltiplos fatores.

[2] As razões-F menores que 1,0 teoricamente também são possíveis – tal resultado indicaria que as diferenças entre médias dos grupos na verdade são *menores* que o esperado por chance. As razões-F muito pequenas refletem uma falha em obter amostras aleatórias independentes. Por exemplo, se a mesma réplica for medida equivocadamente em mais de um grupo de tratamento, a soma dos quadrados entre grupos será pequena (*ver também* Nota de rodapé 6, Capítulo 5).

TABELA 10.4 Tabela de ANOVA para o delineamento de blocos aleatorizados

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)
Entre grupos	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_i-\bar{Y})^2$	$\frac{SQ_{entre\ grupos}}{(a-1)}$
Blocos	$b-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_j-\bar{Y})^2$	$\frac{SQ_{blocos}}{(b-1)}$
Dentro de grupos (resíduo)	$(a-1)(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij}-\bar{Y}_i-Y_j+\bar{Y})^2$	$\frac{SQ_{dentro\ dos\ grupos}}{(a-1)(b-1)}$
Total	$ab-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij}-\bar{Y})^2$	$\frac{SQ_{total}}{(ab-1)}$

Existem de $i = 1$ até a grupos de tratamentos e de $j = 1$ até b blocos, com cada grupo de tratamento representado uma única vez dentro de um bloco. O efeito do tratamento é fixo e o dos blocos, aleatório. Dois testes de hipóteses são possíveis, um para diferenças entre tratamentos e outro para diferenças entre blocos. Os dois testes utilizam o quadrado médio dos resíduos no denominador da razão-F. Como não há replicação dentro do bloco em um delineamento de blocos aleatorizados, não há teste para a interação entre blocos e tratamentos – uma limitação importante desse modelo.

A maioria dos programas estatísticos dará todos esses passos para você. Entretanto, você deveria gastar um tempo para examinar os valores numéricos das razões-F, e se convencer de que eles foram calculados corretamente. Como explicaremos, as configurações-padrão em muitos pacotes estatísticos podem não gerar as razões-F corretas para o seu modelo em particular. Fique atento!

UM BESTIÁRIO DE TABELAS DE ANOVA

Aqui, apresentamos e descrevemos com brevidade as tabelas de ANOVA e os testes de hipóteses para os outros delineamentos de ANOVA, discutidos no Capítulo 7. Também apresentamos um novo delineamento, a ANCOVA, agora que você já aprendeu sobre regressão (no Capítulo 9).

Blocos aleatorizados

No delineamento de blocos aleatorizados, cada conjunto de tratamentos é física ou espacialmente agrupado em um bloco (*ver* Figura 7.6), com cada tratamento representado apenas uma vez em cada bloco. Existem de $a = 1$ até i grupos de tratamentos e de $j = 1$ até b blocos, portanto o tamanho amostral total é de A observações. O modelo em teste é:

Quadrado médio esperado	Razão-F	Valor de *P*
$\sigma^2 + b\sigma_A^2$	$\frac{QM_{entre\ grupos}}{QM_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)$, $(a-1)(b-1)$ graus de liberdade
$\sigma^2 + a\sigma_B^2$	$\frac{QM_{bloco}}{QM_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(b-1)$, $(a-1)(b-1)$ graus de liberdade
σ^2		
σ_Y^2		

$$Y_{ij} = \mu + A_i + B_j + \varepsilon_{ij} \qquad (10.8)$$

Em adição ao termo aleatório de erro ε_{ij} e ao efeito do tratamento A_i, existe agora um efeito de bloco B_j: os valores medidos em alguns blocos são consistentemente maiores ou menores do que em outros blocos, acima e além dos efeitos do tratamento A_i. Note que não há termo de interação incluído para blocos e tratamentos. Tal interação pode evidentemente existir, embora não possamos estimá-la: ela permanece escondida tanto na soma dos quadrados do erro quanto na dos quadrados do tratamento.

A tabela de ANOVA para o delineamento de blocos aleatorizados (Tabela 10.4) contém a soma dos quadrados usual para diferenças entre médias dos tratamentos, mas a dos quadrados para as diferenças entre blocos, que possui $(b-1)$ graus de liberdade. Esta é calculada primeiramente obtendo a média de todos os tratamentos dentro de cada bloco e depois medindo a variação entre eles. A soma dos quadrados do erro agora contém $(a-1)(b-1)$ graus de liberdade. Para a ANOVA de um fator correspondente, com $n = b$ réplicas por tratamento, haverá $(a-1)(b)$ graus de liberdade. Esses números diferem por $(a-1)$ graus de liberdade, que são utilizados para estimar o efeito do bloco.

Duas hipóteses nulas podem ser testadas com o delineamento de blocos aleatorizados. A primeira propõe que não há diferenças entre blocos. A razão-F utilizada para testar essa hipótese é $QM_{entre\ blocos}/QM_{dentro\ dos\ blocos}$. Testar os efeitos do bloco em geral não é de interesse; a principal razão de usar delineamentos em bloco é que esperamos por diferenças entre blocos e queremos ajustar essas

diferenças na nossa comparação de tratamentos. A segunda hipótese nula, a que geralmente estamos mais interessados, é que não há diferenças entre tratamentos. A razão-F utilizada para testar essa hipótese é calculada da maneira usual $QM_{entre\ grupos}/QM_{dentro\ de\ grupos}$. Entretanto, como vimos anteriormente, o quadrado médio dentro de grupos possui menos graus de liberdade que o do erro de um fator simples na ANOVA. A razão é que alguns graus de liberdade originais do erro foram utilizados para estimar o efeito do bloco. Se as diferenças entre os blocos são grandes, a redução na soma dos quadrados entre grupos será considerável e o teste para o efeito do tratamento, mais poderoso, mesmo com reduzidos graus de liberdade. Entretanto, se as diferenças entre blocos são pequenas, a redução na soma dos quadrados entre grupos será diminuta e o teste para o efeito do tratamento menos poderoso. Se os efeitos do bloco são fracos, teremos desperdiçado os $(a - 1)$ graus de liberdade necessários para estimá-los.

ANOVA aninhada

No delineamento aninhado, os dados são organizados hierarquicamente, com uma classe de objetos dentro de outra (*ver* Figura 7.8). Um exemplo familiar é a classificação taxonômica, na qual as espécies são agrupadas dentro de gêneros e estes, agrupados dentro de famílias. O aspecto-chave para reconhecer um delineamento aninhado é que os subgrupos não se repetem em categorias de níveis superiores. Por exemplo, os gêneros de formiga *Myrmica*, *Aphaenogaster* e *Pheidole* ocorrem somente dentro da subfamília de formigas Myrmicinae; eles não são encontrados na subfamília Dolichoderinae ou na subfamília Formicinae. Similarmente, os gêneros de formiga *Formica* e *Camponotus* são encontrados somente na subfamília Formicinae. Os delineamentos aninhados podem lembrar superficialmente os do tipo cruzados ou ortogonais. No entanto, em um delineamento ortogonal adequado, todo o nível de um fator é representado com cada um dos níveis de outro fator. No Capítulo 7, nosso exemplo de delineamento cruzado foi um experimento de adição de nitrogênio e fósforo, no qual cada nível de nitrogênio foi pareado com cada nível de fósforo.

É importante reconhecer as diferenças nesses delineamentos, porque eles requerem diferentes níveis de análise. Embora existam muitas variações de delineamentos aninhados, utilizaremos o mais simples possível, em que um investigador usa duas ou mais subamostras com uma réplica de uma ANOVA de um fator. Desse modo, haverá de $i = 1$ até a níveis de tratamentos, de $j = 1$ até b réplicas dentro de cada nível do tratamento e de $k = 1$ até n subamostras dentro de cada réplica. O tamanho amostral total (para o delineamento balanceado) é $a \times b \times n$. O modelo a ser testado é:

$$Y_{ijk} = \mu + A_i + B_{j(i)} + \varepsilon_{ijk} \qquad (10.9)$$

A_i é o efeito do tratamento e $B_{j(i)}$ é a variação entre réplicas, as quais são aninhadas dentro dos tratamentos. O símbolo $j(i)$ nos lembra que a réplica do nível j é ani-

nhada dentro do nível i do tratamento. Por fim, ε_{ijk} é o termo de erro aleatório, indicando o erro associado à subamostra k, à réplica j e ao tratamento i. Correspondendo a esses três níveis de variação, temos três quadrados médios na tabela de ANOVA: variação entre tratamentos, variação entre réplicas dentro de um tratamento e variação do erro. Esta é calculada para as subamostras dentro de cada réplica (Tabela 10.5).

A característica mais importante na tabela de ANOVA para delineamentos aninhados é a razão-F para o efeito do tratamento. O denominador dessa razão-F é o quadrado médio para as réplicas dentro de um tratamento – e não o usual denominador de variância do erro. A razão é que as subamostras individuais são agrupadas dentro das réplicas, então elas não são independentes uma da outra. Este $QM_{dentro\ de\ grupos}$ é apropriado para testar diferenças entre réplicas dentro de um tratamento, mas *não* para testar diferenças entre tratamentos. O cálculo correto para o efeito do tratamento possui apenas $a(b-1)$ graus de liberdade no denominador, representando a variação independente entre as réplicas.

O resultado do teste de ANOVA aninhada para diferenças entre tratamentos seria algebricamente idêntico a uma ANOVA simples de um fator, em que você primeiro calcula a média das subamostras dentro de uma réplica. Essa ANOVA de um fator também teria $a(b-1)$ graus de liberdade no denominador, os quais correspondem ao número de réplicas verdadeiramente independentes.

Em contraste, se você erroneamente utilizar o $QM_{dentro\ de\ grupos}$ para testar os efeitos dos tratamentos em um delineamento aninhado, você terá $ab(n-1)$ graus de liberdade, que é consideravelmente maior e mais provável de causar uma rejeição incorreta da hipótese nula (um erro Tipo I). A escolha do denominador correto para a razão-F se torna clara quando você examina o quadrado médio esperado na tabela de ANOVA.[3]

A análise que apresentamos aqui representa o delineamento aninhado mais simples possível. Muitos outros delineamentos incluem misturas de fatores aninhados e cruzados, e aqueles de parcelas subdivididas e os de medidas repetidas podem ser interpretados como formas especiais de delineamentos aninhados. Nosso conse-

[3] Alguns autores sugerem que a variância dentro de tratamentos pode ser agrupada sob certas circunstâncias. Suponha que o teste da variação entre réplicas dentro de um tratamento seja não significativo. Então é tentador agrupar a variância e tratar cada subamostra como uma réplica verdadeiramente independente. Com certeza, isso aumentará o poder do teste para o efeito do tratamento, e também a chance de erro Tipo I se realmente houver variação entre as réplicas. O nível crítico para tomar a decisão de agrupar é em geral muito maior (p. ex., $\alpha = 0{,}25$) para reduzir a probabilidade de aceitar incorretamente a hipótese nula de ausência de variação entre as réplicas (um erro Tipo II).

Nosso conselho é não agrupar dados. A escolha de um nível de α para uma decisão de agrupar é razoável, porém arbitrária (Underwood, 1997). Achamos que você deve ficar com a ANOVA aninhada, que verdadeiramente reflete a estrutura dos seus dados. Como destacamos no Capítulo 7, o delineamento aninhado não lhe confere qualquer poder adicional para detectar os efeitos do tratamento e parece perigoso tentar e comprimir os graus de liberdade adicionais para fora da análise depois do fato. Se for logicamente possível, uma outra estratégia é reduzir ou eliminar as subamostragens de seu delineamento e aumentar o número de réplicas verdadeiramente independentes para estimar o efeito do tratamento. Mais uma vez, esta é uma decisão que deve ser resolvida antes de você coletar seus dados.

TABELA 10.5 Tabela de ANOVA para um delineamento aninhado

Fonte	Graus de liberdade (gl)	Soma dos quadrados (SQ)	Quadrado médio (QM)
Entre grupos	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_i-\overline{Y})^2$	$\frac{SQ_{entre\ grupos}}{(a-1)}$
Entre réplicas dentro de grupos	$a(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{j(i)}-\overline{Y}_i)^2$	$\frac{SQ_{réplicas(grupos)}}{a(b-1)}$
Subamostras dentro das réplicas (resíduo)	$ab(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{j(i)})^2$	$\frac{SQ_{subamostras}}{ab(n-1)}$
Total	$abn-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y})^2$	$\frac{SQ_{total}}{(abn-1)}$

Há de $i = 1$ até a grupos de tratamentos, de $j = 1$ até b réplicas aninhadas dentro de cada grupo de tratamento e de $k = 1$ até n subamostras aninhadas dentro de cada réplica. O fator entre grupos é fixo e as réplicas dentro de grupos são tratadas como fatores aleatórios. Note que o teste para o efeito do tratamento usa o quadrado médio entre réplicas, e não o quadrado médio do resíduo. O quadrado médio dos resíduos (subamostras) é um denominador no teste da variação entre réplicas dentro de tratamentos. Um erro comum em análises de delineamentos aninhados é tratar as subamostras como independentes (o que elas não são) e usar o quadrado médio dos resíduos no denominador da razão-F para testar o efeito dos tratamentos.

lho é evitar delineamentos complicados com alguns fatores aninhados e cruzados. Em alguns casos, nem mesmo será possível construir um modelo de ANOVA válido para esses delineamentos. Se seus dados estão organizados em um delineamento aninhado complicado, você sempre poderá analisar as médias das subamostras não independentes — com frequência isso irá colapsar seu delineamento em um modelo mais simples. Como discutimos no Capítulo 7, a média das subamostras reduz o seu tamanho amostral aparente, mas preserva os graus de liberdade verdadeiros e a replicação independente, necessária para testes de hipóteses válidos.

ANOVA de dois fatores

Retornamos ao exemplo de delineamento com dois fatores do Capítulo 7: um estudo do recrutamento de cracas onde um fator é o substrato (3 níveis: cimento, ardósia e granito) e o segundo fator é predação (4 níveis: sem manipulação, controle, exclusão de predador e inclusão de predador). Há de $i = 1$ até a níveis do primeiro fator, de $j = 1$ até b níveis do segundo fator e de $k = 1$ até n réplicas de cada combinação ij única dos tratamentos. Existem ab combinações únicas de tratamentos e um total de $a \times b \times n$ réplicas. No exemplo do Capítulo 7, $a = 3$ níveis de substratos, $b = 4$ níveis de predação e $n = 10$ réplicas por combinação de

Quadrado médio esperado	Razão-F	Valor de *P*
$\sigma^2 + bn\sigma^2_A + n\sigma^2_{B(A)}$	$\frac{QM_{entre\ grupos}}{QM_{entre\ réplicas\ (grupos)}}$	Cauda da distribuição F com $(a-1), a(b-1)$ graus de liberdade
$\sigma^2 + n\sigma^2_{B(A)}$	$\frac{QM_{entre\ réplicas\ (grupos)}}{QM_{subamostras}}$	Cauda da distribuição F com $a(b-1), ab(n-1)$ graus de liberdade
σ^2		
σ^2_Y		

tratamentos. Há $ab = (3)(4) = 12$ combinações únicas de tratamentos e um tamanho amostral total de $a \times b \times n = (3)(4)(10) = 120$ (*ver* Figura 7.9).

Em vez de um único quadrado médio para representar o efeito do tratamento como na ANOVA de um fator, há, agora, três quadrados médios para os fatores de tratamentos. Há uma soma dos quadrados e um quadrado médio para cada **efeito principal**, ou grupo de tratamento: um para o substrato com $(a-1) = 2$ graus de liberdade e um para predação com $(b-1) = 3$ graus de liberdade. Essas somas dos quadrados são usadas para testar diferenças nas médias de cada fator, exatamente como na ANOVA de um fator.

Entretanto, uma diferença sutil é que a soma dos quadrados, associada ao efeito principal do substrato na ANOVA de dois fatores, é calculada *tirando a média* de todos os níveis de predação. De forma similar, a soma dos quadrados associada ao efeito principal de predação é calculada tirando a média de todos os níveis de substrato. Em contraste, o efeito principal da ANOVA de um fator simplesmente tira a média das réplicas dentro de cada tratamento, pois não há um segundo fator presente no delineamento experimental.[4]

[4] É claro, este segundo fator ainda está presente na natureza, mas ele não foi incorporado ao delineamento. O delineamento de um fator pode restringir o experimento a apenas um nível do outro fator. Por exemplo, um delineamento de um fator para o experimento de predação pode ser conduzido somente em substratos naturais. Como alternativa, o delineamento de um fator pode não controlar explicitamente o segundo fator. Entretanto, se o delineamento incluir aleatorização e replicação adequadas (*ver* Capítulo 6), o segundo fator apenas contribuirá para a variação do resíduo. Por exemplo, o delineamento de um fator para o experimento de substrato ignora a predação, mas se as réplicas forem independentes entre si e colocadas de forma aleatória, o efeito da predação será um componente não explicado da variação residual.

Em adição aos quadrados médios dos dois efeitos principais, há um terceiro efeito que é estimado na ANOVA de dois fatores: a interação entre ambos. O **efeito de interação** mede diferenças nas médias dos grupos de tratamentos que não podem ser preditas em bases aditivas dos dois efeitos principais. Quantificar os efeitos de interação de dois ou mais tratamentos em geral é a razão-chave para conduzir um experimento multifatorial ou um estudo amostral. Posteriormente, neste capítulo, estudaremos o efeito de interação em mais detalhes. Por agora, apenas notaremos que o efeito de interação possui $(a-1)(b-1)$ graus de liberdade. No exemplo das cracas, esse efeito possui $(3-1)(4-1)=6$ graus de liberdade.

Os dois efeitos principais e o termo de interação aparecem como elementos em nosso modelo:

$$Y_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk} \qquad \textbf{(10.10)}$$

Seguindo nosso procedimento para construir razões-F, cada um dos quadrados médios correspondentes será usado no numerador e o termo do erro será

TABELA 10.5 Tabela de ANOVA para um delineamento aninhado

Fonte	Graus de liberdade (gl)	Soma dos quadrados (SQ)	Quadrado médio (QM)
Fator A	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_i-\overline{Y})^2$	$\frac{SQ_A}{(a-1)}$
Fator B	$b-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_j-\overline{Y})^2$	$\frac{SQ_B}{(b-1)}$
Interação $(A\times B)$	$(a-1)(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{ij}-\overline{Y}_i-\overline{Y}_j+\overline{Y})^2$	$\frac{SQ_{AB}}{(a-1)(b-1)}$
Dentro de grupos (resíduo)	$ab(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{ij})^2$	$\frac{SQ_{dentro\ de\ grupos}}{ab(n-1)}$
Total	$abn-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y})^2$	$\frac{SQ_{total}}{(abn-1)}$

Há de $i=1$ até a níveis do Fator A, de $j=1$ até b níveis do Fator B e n réplicas de cada combinação única de tratamentos ij. Como os Fatores A e B são fixos, os efeitos principais e os da interação são todos testados contra o quadrado médio dos resíduos. Este modelo é o resultado padrão de quase todos os pacotes estatísticos, embora em muitos delineamentos os Fatores A ou B possam ser aleatórios, e não fixos.

sempre utilizado no denominador (Tabela 10.6). Para o delineamento de dois fatores, os graus de liberdade do termo do erro no denominador será $ab(n-1) = (4)(3)(9) = 108$ graus de liberdade. Contudo, esse procedimento padrão para a ANOVA de dois fatores é válido apenas se ambos forem conhecidos por **efeitos fixos**. Se forem de **efeitos aleatórios**, o quadrado médio esperado muda e temos que construir razões-F diferentemente. Primeiro, trabalharemos com tabelas típicas de ANOVA para fatores fixos e, então, retornaremos a esse problema posteriormente, quando discutirmos fatores aleatórios e fixos.

Por fim, é instrutivo contrastar o delineamento de dois fatores para esses dados (2 fatores com 4 e 3 níveis de tratamento, respectivamente) com o delineamento de um fator correspondente (1 fator com 12 níveis de tratamento). No delineamento de um fator, o termo de erro possui $a(n-1) = 12\ (10-1) = 108$ graus de liberdade. Para o delineamento de dois fatores, os graus de liberdade do denominador para o termo de erro são $ab(n-1) = (4)(3)(9) = 108$ graus de liberdade. Portanto, os cálculos dos graus de liberdade e dos quadrados médios são idênticos, independentemente se analisamos os dados como uma configuração de um fator ou como uma configuração de dois fatores adequada.

Quadrado médio esperado	Razão-F	Valor de *P*
$\sigma^2 + nb\sigma^2_A$	$\frac{QM_A}{QM_{dentro\ dos\ grupos}}$	Cauda da distribuição F com $(a-1), ab(n-1)$ graus de liberdade
$\sigma^2 + na\sigma^2_B$	$\frac{QM_B}{QM_{dentro\ dos\ grupos}}$	Cauda da distribuição F com $(b-1), ab(n-1)$ graus de liberdade
$\sigma^2 + n\sigma^2_{AB}$	$\frac{QM_{AB}}{QM_{dentro\ dos\ grupos}}$	Cauda da distribuição F com $(a-1)(b-1), ab(n-1)$ graus de liberdade
σ^2		
σ^2_Y		

Compare com cuidado os graus de liberdade para os efeitos do tratamento nesses dois modelos. Para o delineamento de dois fatores, se você adicionar os graus de liberdade dos dois efeitos principais e da interação terá: 2 (efeito principal de substrato) + 3 (efeito principal de predação) + 6 (interação predação × substrato) = 11. Esse é o mesmo número de graus de liberdade que tínhamos no delineamento de um fator com 12 níveis de tratamento. Além disso, você poderia encontrar que a soma dos quadrados desses dois termos também possui o mesmo valor. Efetivamente fizemos a partição dos graus de liberdade dos tratamentos da ANOVA de um fator em componentes que refletem a estrutura lógica de um esquema para dois fatores.

ANOVA para delineamentos com três fatores e para *n*-fatores

Em teoria, o delineamento de dois fatores pode ser estendido para qualquer número de fatores. Cada fator possui níveis diferentes de fatores e todos os tratamentos são completamente cruzados. Cada nível de um tratamento é aplicado com cada nível de todos os outros tratamentos, até que todas as combinações sejam representadas. Por exemplo, um experimento de três fatores que manipula a presença ou ausência de herbívoros, carnívoros e predadores (Tabela 7.3) possui dois níveis para cada um dos três fatores. No modelo de três fatores, há uma média geral (μ), três efeitos principais (A, B, C), três pares de interações (AB, AC, BC), um termo de interação de três fatores (ABC) e um termo do erro (ε). O modelo é:

$$Y_{ijkl} = \mu + A_i + B_j + C_k + AB_{ij} + AC_{ik} + BC_{jk} + ABC_{ijk} + \varepsilon_{ijkl} \quad (10.11)$$

Os efeitos principais de herbívoro, carnívoro e predador possuem cada um ($a - 1$) graus de liberdade (1 neste exemplo). Os termos de interação em pares representam os efeitos não aditivos de cada par de fatores tróficos possível. Cada um desses termos de interação possui $(a - 1)(b - 1)$ graus de liberdade (esta expressão também é igual a 1 neste exemplo). Por fim, há um único termo de interação de três fatores com $(a - 1)(b - 1)(c - 1)$ graus de liberdade (coincidentemente também é igual a 1 neste exemplo). Como anteriormente (com um modelo de efeitos fixos), todos os efeitos principais e os termos de interação são testados utilizando o termo de erro como quadrado médio do denominador (Tabela 10.7).

ANOVA de parcelas subdivididas

No delineamento de parcelas subdivididas (*split-plot*), os tratamentos de um fator são agrupados espacialmente juntos, como em um delineamento de blocos aleatorizados. Um segundo tratamento é então aplicado ao bloco ou à parcela inteira (*ver* Figura 7.12). No exemplo das cracas, no Capítulo 7, o fator completo da parcela é a predação, pois o bloco inteiro é cercado ou manipulado. O fator de

dentro da parcela é o substrato, pois cada tipo de substrato é representado dentro de cada bloco. O modelo é:

$$Y_{ijk} = \mu + A_i + B_{j(i)} + C_k + AC_{ik} + CB_{kj(i)} [+ \varepsilon_{ijkl}] \qquad (10.12)$$

O tratamento completo das parcelas é A, as parcelas diferentes (aninhadas dentro do fator A) são $B_{j(i)}$, o fator dentro das parcelas é C_k e o termo de erro é ε_{ijkl}. Apresentamos esse termo de erro entre colchetes para elucidar o modelo completo, mas ele não pode ser isolado nesse modelo, pois não há replicação do fator C dentro dos blocos (cada nível de C é representado apenas uma vez dentro de um bloco).

Dois termos de erro diferentes são usados para testar hipóteses no delineamento de parcelas subdivididas. Para testar os efeitos do tratamento completo A, utilizamos o quadrado médio dos blocos $B_{j(i)}$ como denominador da razão-F. Isso porque os blocos inteiros servem como réplicas independentes em relação aos tratamentos da parcela completa.

O tratamento C dentro da parcela é testado contra o termo de interação $C \times B$, assim como a interação $A \times C$ (Tabela 10.8). Como na ANOVA de dois fatores padrão, este modelo gera razões-F e testes de hipóteses para os efeitos principais A, C e para a interação entre eles ($A \times C$). Entretanto, como o termo do erro residual (ε_{ijkl}) não poder ser completamente isolado, o modelo de parcelas subdivididas assume que não há interação entre o fator C e as subparcelas ($CB_{kj(i)} = 0$; *ver* Underwood, 1997, página 393*ff*, para uma discussão completa).

ANOVA de medidas repetidas

O delineamento de medidas repetidas é aquele no qual múltiplas observações são tomadas em um único indivíduo ou réplica. Se há muita variabilidade de uma réplica para a próxima, esta técnica controla essa fonte de variação. Contudo, as observações repetidas em um mesmo indivíduo ou réplica não são estatisticamente independentes entre si, devendo-se tomar cuidado para garantir que as análises reflitam essa estrutura de dependência nos dados.

Dois tipos de delineamentos são incluídos na análise de medidas repetidas. No primeiro, única réplica é exposta a diferentes tratamentos experimentais, cada um aplicado em tempos diferentes e em ordem aleatorizada. Por exemplo, plantas individuais podem ser expostas a uma série de diferentes concentrações de CO_2. Em cada uma, as taxas fotossintéticas são medidas (p. ex., Potvin et al., 1986). Esse delineamento também pode ser utilizado quando as réplicas são amostradas repetidas vezes, mas nenhuma manipulação é aplicada. Neste caso, os tratamentos correspondem apenas aos efeitos do tempo.

Para ambas as variações, o modelo é:

$$Y_{ij} = \mu + A_i + B_j + \varepsilon_{ij} \qquad (10.13)$$

TABELA 10.7 Tabela de ANOVA para um delineamento fatorial com 3 fatores

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)
Fator *A*	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(\overline{Y}_i-\overline{Y})^2$
Fator *B*	$b-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(\overline{Y}_j-\overline{Y})^2$
Fator *C*	$c-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(\overline{Y}_k-\overline{Y})^2$
Interação $A\times B$	$(a-1)(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(\overline{Y}_{ij}-\overline{Y}_i-\overline{Y}_j+\overline{Y})^2$
Interação $A\times C$	$(a-1)(c-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(\overline{Y}_{ik}-\overline{Y}_i-\overline{Y}_k+\overline{Y})^2$
Interação $B\times C$	$(b-1)(c-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(\overline{Y}_{jk}-\overline{Y}_j-\overline{Y}_k+\overline{Y})^2$
Interação $A\times B\times C$	$(a-1)(b-1)(c-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}\left(\overline{Y}_{ijk}-\overline{Y}_{ij}-\overline{Y}_{ik}-\overline{Y}_{jk}+\overline{Y}_i+\overline{Y}_j+\overline{Y}_k-\overline{Y}\right)^2$
Dentro de grupos (resíduo)	$abc(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(Y_{ijkl}-\overline{Y}_{ijk})^2$
Total	$abcn-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n}(Y_{ijkl}-\overline{Y})^2$

Há de $i=1$ até a níveis do Fator *A*, de $j=1$ até b níveis do Fator *B*, de $k=1$ até c níveis do Fator *C* e n réplicas de cada combinação única de tratamento ijk. Os fatores *A*, *B* e *C* são fixos. Este modelo gera três efeitos principais (*A*, *B*, *C*), três termos de interação par a par (*AB*, *AC*, *BC*) e um termo de interação com três fatores (*ABC*). Como os Fatores *A*, *B* e *C* são todos fixos, todos os efeitos principais e termos de interação são testados contra o quadrado médio do resíduo. Com o intuito de analisar esse tipo de delineamento, todas as possíveis combinações de tratamentos dos diferentes fatores devem ser estabelecidas com replicação para cada uma delas.

Quadrado médio (QM)	Quadrado médio esperado	Razão-F	Valor de P
$\frac{SQ_A}{(a-1)}$	$\sigma^2 + bcn\sigma_A^2$	$\frac{SQ_A}{SQ_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)$, $ab(n-1)$ graus de liberdade
$\frac{SQ_B}{(b-1)}$	$\sigma^2 + acn\sigma_B^2$	$\frac{SQ_B}{SQ_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(b-1)$, $ab(n-1)$ graus de liberdade
$\frac{SQ_C}{(c-1)}$	$\sigma^2 + abn\sigma_C^2$	$\frac{SQ_C}{SQ_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(c-1)$, $abc(n-1)$ graus de liberdade
$\frac{SQ_{AB}}{(a-1)(b-1)}$	$\sigma^2 + cn\sigma_{AB}^2$	$\frac{SQ_{AB}}{SQ_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)$ $(b-1)$, $abc(n-1)$ graus de liberdade
$\frac{SQ_{AC}}{(a-1)(c-1)}$	$\sigma^2 + bn\sigma_{AC}^2$	$\frac{SQ_{AC}}{SQ_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)$ $(c-1)$, $abc(n-1)$ graus de liberdade
$\frac{SQ_{BC}}{(b-1)(c-1)}$	$\sigma^2 + an\sigma_{BC}^2$	$\frac{SQ_{BC}}{SQ_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(b-1)$ $(c-1)$, $abc(n-1)$ graus de liberdade
$\frac{SQ_{ABC}}{(a-1)(b-1)(c-1)}$	$\sigma^2 + n\sigma_{ABC}^2$	$\frac{SQ_{ABC}}{SQ_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)$ $(b-1)(c-1)$, $abc(n-1)$ graus de liberdade
$\frac{SQ_{dentro\ de\ grupos}}{abc(n-1)}$	σ^2		
$\frac{SQ_{total}}{(abcn-1)}$	σ_Y^2		

onde há de $i = 1$ até a tratamentos (ou tempos) e de $j = 1$ até n réplicas (ou indivíduos). Se você comparar as Equações 10.13 e 10.8, verá que o modelo de medidas repetidas (e a tabela de ANOVA correspondente) é o mesmo que utilizamos para um delineamento de blocos aleatorizados (Tabela 10.4). Cada indivíduo serve como seu próprio bloco e os graus de liberdade do erro são ajustados de maneira adequada. Como no delineamento de blocos aleatorizados, esta análise assume que não há interação entre réplicas (indivíduos) e tratamentos – apenas diferenças aditivas sim-

TABELA 10.8 Tabela de ANOVA para delineamento de parcelas subdivididas

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)
Fator *A* (tratamento parcela completa)	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\overline{Y}_i-\overline{Y})^2$	$\frac{SQ_A}{(a-1)}$
Fator *B*(*A*) (parcelas aninhadas dentro de *A*)	$a(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\overline{Y}_j-\overline{Y})^2$	$\frac{SQ_B}{(b-1)}$
Fator *C* (tratamento dentro das parcelas)	$(c-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\overline{Y}_k-\overline{Y})^2$	$\frac{SQ_C}{(c-1)}$
Interação $A \times C$	$(a-1)(c-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(\overline{Y}_{ik}-\overline{Y}_i-\overline{Y}_k+\overline{Y})^2$	$\frac{SQ_{AC}}{(a-1)(c-1)}$
Interação $B(A) \times C$	$a(b-1)(c-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(Y_{ijk}-\overline{Y}_{ik})^2$	$\frac{SQ_{BC}}{a(b-1)(c-1)}$
Total	$abc-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(Y_{ijk}-\overline{Y}_.)^2$	$\frac{SQ_{total}}{(abc-1)}$

Há de $i = 1$ até a níveis do Fator *A* (aplicado às parcelas inteiras), de $j = 1$ até b parcelas aninhadas dentro do Fator *A* e de $k = 1$ até c níveis do Fator *C* (aplicado dentro das parcelas). Os Fatores *A* e *C* são fixos, o Fator *B* (parcelas) é aleatório. Neste delineamento, a razão-F para o fator *A* utiliza o fator parcela (*B*, aninhado dentro de *A*) como o termo de erro para o denominador da razão-F, pois são as parcelas que são as réplicas independentes para o Fator *A*. O teste da razão-F para o Fator *C* (o fator dentro da parcela) e para a interação $A \times C$ utilizam a interação *B* (parcela) × *C* como denominador. Este modelo assume que não há interação entre parcelas e o Fator *A*, porque essa interação não pode ser estimada no delineamento de parcelas subdivididas.

ples entre indivíduos em suas respostas aos tratamentos. Além disso, os tratamentos precisam ser aplicados em ordem aleatória e deve ter passado tempo suficiente entre as aplicações dos tratamentos para garantir que as respostas sejam independentes. Temos de assumir que não há efeitos carregados de um tratamento para o próximo, como a acumulação de nutrientes, alteração de hábitat, adaptação fisiológica ou (em estudos comportamentais) habituação ou respostas aprendidas. Note, também, que a ordem do tratamento é confundida com o tempo: como os tratamentos não são aplicados simultaneamente, um indivíduo recebe certos tratamentos no início do período experimental e outros em períodos posteriores. Esta é outra razão do porque é essencial aleatorizar a ordem de aplicação dos tratamentos.

No segundo tipo de delineamento de medidas repetidas, diferentes tratamentos são aplicados, porém cada réplica (ou indivíduo) recebe um único tratamento,

Quadrado médio esperado	Razão-F	Valor de *P*
$\sigma^2 + c\sigma^2_B + bc\sigma^2_A$	$\frac{QM_A}{QM_B}$	Cauda da distribuição F com $(a-1), a(b-1)$ graus de liberdade
$\sigma^2 + c\sigma^2_B$		
$\sigma^2 + \sigma^2_{BC} + ba\sigma^2_C$	$\frac{QM_C}{QM_{B(A)xC}}$	Cauda da distribuição F com $(c-1)$, $a(b-1)(c-1)$ graus de liberdade
$\sigma^2 + \sigma^2_{BC} + b\sigma^2_{AC}$	$\frac{QM_{AC}}{QM_{B(A)xC}}$	Cauda da distribuição F com $(a-1)$ $(c-1), a(b-1)(c-1)$ graus de liberdade
$\sigma^2 + \sigma^2_{BC}$		
σ^2_Y		

que depois é medida em tempos diferentes. Você pode pensar nesse delineamento como uma extensão da ANOVA de um fator, em vez de fazer uma única observação por réplica, aplicamos o tratamento e fazemos observações múltiplas em cada réplica em diferentes tempos. O modelo é:

$$Y_{ijk} = \mu + A_i + B_{j(i)} + C_k + AC_{ik} + CB_{kj(i)} \qquad \textbf{(10.14)}$$

Neste modelo, há de $i = 1$ até a tratamentos, de $j = 1$ até b réplicas (ou indivíduos) aninhados dentro dos tratamentos e de $c = 1$ até k vezes (tempos). Se você comparar a Equação 10.14 com a 10.12 verá que esse modelo de medidas repetidas é estruturalmente equivalente ao de parcelas subdivididas. No delineamento de medidas repetidas, cada réplica (ou indivíduo) é equivalente a uma parcela inteira. O fator de parcela completa é simplesmente o tratamento que foi aplicado no delineamento de um fator. O fator dentro da parcela é o tempo, com cada período amostral equivalente a um nível de tratamento diferente. Na terminologia do delineamento de medidas repetidas, o efeito do tratamento da parcela completa corresponde ao entre temas, e os efeitos do tratamento dentro das parcelas correspondem ao existente dentro de temas.

Como no delineamento de parcelas subdivididas, temos de assumir que não há interação entre tempo e réplica: em outras palavras, o perfil da tendência temporal da resposta é o mesmo para todas as réplicas dentro de um dado tratamento. Como cada indivíduo é, por definição, único, não há nenhuma amostragem adicional que podemos realizar para estimar esse termo de interação. Uma última variante do delineamento de medidas repetidas é aquela na qual as medidas de "tempo" são feitas antes e depois da aplicação do tratamento. Nesse caso, temos a forma de um delineamento ADCI (Antes-Depois-Controle-Impacto), com medidas antes e depois em parcelas de controle e de impacto (*ver* Capítulo 7).

Como notamos no Capítulo 7, todos esses delineamentos assumem circularidade estatística – a variância da diferença entre quaisquer pares de tempo é a mesma. Tal pressuposto raras vezes é cumprido com dados de séries temporais e outros tipos de análises podem ser mais apropriadas para incorporar a variação temporal.[5] Como na análise de parcelas subdivididas, você deve evitar o erro comum de encarar esses dados como um delineamento apenas de dois fatores. As observações repetidas coletadas em uma única réplica não são estatisticamente independentes, portanto o quadrado médio da variação entre réplicas dentro de um tratamento é o denominador correto para usar no seu teste do efeito dos tratamentos.

ANCOVA

Agora que você entende como usar a regressão linear para variáveis contínuas (Capítulo 9), podemos desenvolver um novo modelo, híbrido de regressão e análise de variância. A **ANCOVA (análise de covariância)** é utilizada para delineamentos de ANOVA, na qual uma variável contínua adicional (a **covariável**) é medida para cada réplica. A hipótese é de que a covariável também contribui para a alteração na variável resposta. Se a covariável não tiver sido medida, aquela fonte de variação acumularia como erro puro no resíduo. Na análise de covariância, podemos remover estatisticamente essa fonte de variação dos resíduos. Se a covariável tiver um efeito importante, o tamanho do resíduo é muito menor e nosso teste para as diferenças dos tratamentos será muito mais poderoso. Uma forma útil de pensar sobre a ANCOVA é que ela é uma análise de variância realizada com os resíduos de uma regressão da variável resposta pela covariável.

Retornando ao exemplo das cracas do Capítulo 7, sabemos que o seu estabelecimento é diferente em superfícies horizontais *versus* verticais e suspeitamos que a variação no ângulo em que colocamos cada réplica pode contribuir para

[5] Discutimos as análises univariadas para dados de medidas repetidas. Uma estratégia diferente é analisar os dados de medidas repetidas como de variância multivariada (MANOVA; *ver* Capítulo 12), na qual o vetor resposta é um conjunto de medidas repetidas para cada réplica. A MANOVA possui pressupostos diferentes daqueles das análises univariadas e ela pode ser menos poderosa e requerer mais replicação. *Ver* Gurevitch & Chester (1986) e Potvin et al. (1990) para uma discussão adicional.

o padrão em nossos dados. É claro que tentaríamos colocar nossos tratamentos de substratos em uma orientação aproximadamente horizontal, mas é provável que haja variação de uma réplica para a outra. Por isso, para cada uma, medimos o ângulo de orientação em uma escala contínua que vai de 0 até 90 graus.[6] Essa variável contínua será utilizada como covariável nas análises.

O modelo é:

$$Y_{ij} = \mu + A_i + \beta_i(X_{ij} - \overline{X}_i) + \varepsilon_{ij} \qquad \textbf{(10.15)}$$

onde A_i é o efeito do tratamento (de $i = 1$ até a tratamentos), X_{ij} é a covariável medida para a observação Y_{ij}, $\overline{X}_i$ é o valor médio da covariável para o grupo de tratamento i e ε_{ij} é (como de costume) o termo de erro. Note que este modelo contém um termo A_i para o efeito do tratamento (como na ANOVA) e um termo de inclinação β_i para o efeito da covariável (como na regressão). A Equação 10.15 representa o caso mais complexo. Isso implica que cada grupo de tratamento é descrito por sua própria linha de regressão, cada uma com inclinação e intercepto distintos.

Um número de modelos simples está incorporado na Equação 10.15. Por exemplo, suponha que todos os grupos de tratamentos tenham uma inclinação de regressão em comum. Neste caso, podemos utilizar o mesmo termo de inclinação (β_C) para cada grupo de tratamento (β_i):

$$Y_{ij} = \mu + A_i + \beta_C(X_{ij} - \overline{X}_i) + \varepsilon_{ij} \qquad \textbf{(10.16)}$$

Se os termos de inclinação representando o efeito da covariável não diferirem de zero, o modelo reduz a:

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}$$

que é apenas uma ANOVA de um fator (Equação 10.6).

Inversamente, se não há efeito do tratamento e se não houver interação entre o tratamento e a covariável, o modelo reduz a uma regressão linear pela covariável:

[6] Seja cuidadoso quando utilizar dados angulares de qualquer tipo. O exemplo das cracas que descrevemos aqui pode ser analisado com estatísticas convencionais, mas se os dados angulares forem medidos em uma escala circular completa, haverá problema. Por exemplo, em um estudo de atividade de formigas, suponha que medimos a atividade de forrageamento diária em uma escala de 0 a 24 horas. Se o pico de atividade é a meia-noite, poderíamos ter registrado duas observações de forrageamento às 23h (= 11:00 pm) e a 1h (= 1:00 am). Mas se você tirar a média desses dois números, a média será meio-dia (12h) e não meia-noite (0h)! Como a escala dessas medidas é circular, em vez de linear, não podemos utilizar as fórmulas-padrão para calcular nem mesmo medidas básicas, como médias e variâncias. Um conjunto inteiro de procedimentos diferentes é necessário para essas estatísticas circulares. *Ver* Fischer (1993) para uma introdução a estatísticas circulares em biologia.

$$Y_{ij} = \mu + \beta_T(X_{ij} - \overline{X}) + \varepsilon_{ij} \qquad (10.17)$$

A inclinação neste modelo (β_T) é ajustada para todos os dados e as atribuições de tratamentos são ignoradas.

A análise de covariância geralmente envolve primeiro um teste das diferenças entre os β_i. Esse teste pergunta se as inclinações são heterogêneas para o grupo como um todo. Se positivo, o efeito do tratamento depende do valor da covariável (a variável X) e utilizamos o modelo completo para as análises (Equação 10.15). Em casos extremos, se as linhas das regressões se cruzarem, a ordem dos valores esperados para cada tratamento irá diferir entre valores pequenos e grandes da covariável. Essa "interação" é similar a um termo de interação entre duas variáveis categóricas em uma ANOVA fatorial: as diferenças entre as médias dos tratamentos do Fator *A* dependem do nível do Fator *B* (e vice-versa).

Se o teste da heterogeneidade das inclinações não for significativo, uma inclinação em comum β_C pode ser ajustada para todos os tratamentos (Equação 10.16) e as médias ajustadas podem ser calculadas. A **média ajustada** é o valor esperado do grupo de tratamento que seria encontrado se todos os grupos tiverem o mesmo escore médio da covariável:

$$\overline{Y}_i = \mu + A_i + \beta_C \overline{X} \qquad (10.18)$$

Idealmente, em uma análise de covariância, a medida da covariável para os diferentes grupos de tratamentos abrangerá a mesma amplitude de valores e não irá diferir em suas médias (Figura 10.1A). Neste caso, a ANCOVA faz sentido porque os dados nos grupos de tratamentos são comparados à mesma amplitude de valores da covariável. Essa distribuição dos escores da covariável é exatamente o que esperaríamos se métodos de aleatorização apropriados fossem utilizados para determinar o arranjo espacial das réplicas e as atribuições dos tratamentos. Mas suponha que no experimento das craca̲s você coloque inadvertidamente o substrato de cimento em uma área costeira plana, o de granito em uma área íngreme e o de ardósia em uma área com inclinação intermediária (Figura 10.1B). Agora, a ANCOVA está em terreno movediço, pois os tratamentos estão confundidos com as diferenças na covariável. Com relação às estatísticas, ainda podemos proceder com a ANCOVA, mas teremos que extrapolar a linha da regressão para além da amplitude dos dados de cada tratamento e assumir (bem como ter esperança) que a mesma relação linear ainda se mantenha. Se a linearidade se mantém, a ANCOVA irá estatisticamente fazer a partição da variância, que é associada à covariável. Porém, ela não irá remediar o fato de que o delineamento está inerentemente confundido. Como temos enfatizado ao longo desse texto, a aleatorização é a melhor proteção contra surpresas como a da Figura 10.1B.

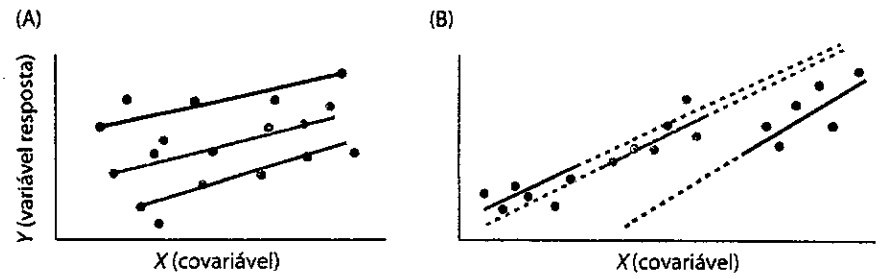


Figura 10.1 Conjunto de dados seguros e arriscados para uma análise de covariância. A covariável é apresentada no eixo-*x* e a variável resposta no eixo-*y*. As três cores representam três grupos de tratamentos. Cada ponto é uma réplica independente. (A) Neste caso, as medidas da covariável ocorrem na mesma amplitude para cada grupo de tratamento. As comparações da ANCOVA são seguras, porque as relações de regressão não precisam ser extrapoladas. (B) Neste caso, as medidas da covariável para cada um dos três grupos de tratamentos não se sobrepõem. O delineamento está essencialmente confundido, pois o tratamento verde está associado somente aos valores pequenos da covariável e os tratamentos pretos somente aos valores grandes. A análise assume que as relações de regressão permanecem lineares quando elas são extrapoladas para além da amplitude observada dos dados (indicado pelas linhas de regressão tracejadas). A análise estatística é a mesma em ambos os casos, porém as conclusões são muito menos confiáveis para os dados em (B).

FATORES ALEATÓRIOS *VERSUS* FATORES FIXOS EM ANOVA

Os testes que apresentamos até agora são os tipos de resultados que você obteria em uma análise padrão com qualquer programa de computador. Infelizmente, para delineamentos com dois fatores ou mais complicados, essas análises podem não ser corretas.

Um ponto importante que surge em delineamentos multifatoriais é quando cada fator deve ser analisado como um **fator fixo** ou como um **fator aleatório.** No primeiro caso, os diferentes níveis de tratamento utilizados são os únicos de interesse, e as inferências serão restritas àqueles níveis em particular. No segundo, os níveis de tratamento em particular representam uma amostra aleatória de todos os níveis possíveis que podem ter sido estabelecidos; pretende-se que as inferências sejam válidas não apenas para os níveis de tratamento testados, mas para outros que não foram incluídos no delineamento. Em uma **análise de modelo misto**, alguns dos fatores são fixos e outros são aleatórios. O resultado padrão em quase todos os programas estatísticos é para uma ANOVA com fatores fixos.

Por que essa distinção entre fatores fixos e aleatórios é importante? A razão é que os valores dos quadrados médios esperados são diferentes em ANOVAs com fatores fixos ou aleatórios. Consequentemente, as razões-F utilizadas para testar hipóteses acerca de fatores aleatórios e fixos serão diferentes. Por exemplo, em uma ANOVA com dois fatores fixos, tanto os efeitos principais quanto os termos

de interação são testados contra o quadrado médio dos resíduos (Tabela 10.6). Mas em um modelo de efeitos aleatórios, os efeitos principais são testados utilizando o quadrado médio da interação no denominador da razão-F e não o quadrado médio dos resíduos. Tanto os quadrados médios quanto os graus de liberdade são muito diferentes para esses dois termos, o que significa que seus valores de significância mudarão completamente. Em um modelo misto, o Fator *A* é fixo e o Fator *B* é aleatório. Neste caso, o Fator *A* é testado contra o quadrado médio da interação, mas o Fator *B* é testado contra o quadrado médio dos resíduos. As Tabelas 10.9 e 10.10 fornecem os quadrados médios esperados e as razões-F apropriadas tanto para modelos de efeitos aleatórios quanto para modelos mistos para uma ANOVA de dois fatores. Note que ANOVAs de fatores fixos, de fatores aleatórios e de modelos mistos se aplicam apenas para modelos com dois ou mais fatores. ANOVAs de um fator são calculadas da mesma forma, independente se o fator do tratamento é fixo ou aleatório.

O cálculo correto das razões-F em ANOVAs de modelos com fatores aleatórios ou mistos é muito importante, mas tão importante quanto é o fato de que esses diferentes modelos prescrevem diferentes estratégias amostrais. Nos de efeitos fixos, nosso objetivo deve ser o de replicar o máximo que conseguirmos para cada combinação de tratamentos. Todas as réplicas extras nos darão mais

TABELA 10.9 Tabela de ANOVA para uma ANOVA de dois fatores com efeitos aleatórios

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)
Fator *A*	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_i-\bar{Y})^2$	$\frac{SQ_A}{(a-1)}$
Fator *B*	$b-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_j-\bar{Y})^2$	$\frac{SQ_B}{(b-1)}$
Interação ($A \times B$)	$(a-1)(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{ij}-\bar{Y}_i-\bar{Y}_j+\bar{Y})^2$	$\frac{SQ_{AB}}{(a-1)(b-1)}$
Dentro de grupos (resíduos)	$ab(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y}_{ij})^2$	$\frac{SQ_{dentro\ de\ grupos}}{ab(n-1)}$
Total	$abn-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y})^2$	$\frac{SQ_{total}}{(abn-1)}$

Há de $i = 1$ até a níveis do Fator *A*, de $j = 1$ até b níveis do Fator B e n réplicas para cada combinação única de tratamentos ij. A razão-F para ambos os efeitos principais é testada utilizando o quadrado médio da interação no denominador. Em contraste, o quadrado médio dos resíduos é utilizado no denominador da razão-F padrão (Tabela 10.6), na qual ambos os tratamentos são fatores fixos.

graus de liberdade para o quadrado médio dos resíduos, o que irá aumentar nosso poder estatístico. Mas, para modelos de efeitos aleatórios ou mistos, devemos utilizar uma estratégia diferente. Em vez de tentar aumentar a replicação dentro de cada nível de tratamento, devíamos procurar aumentar o número de níveis de tratamento que estabelecemos e não nos preocupar muito com a replicação dentro de cada nível. O número de níveis de tratamento determina o número de graus de liberdade do quadrado médio da interação, e é onde queremos aumentar a amostragem. No pior cenário, se rodarmos uma ANOVA de dois fatores aleatórios com apenas dois níveis de tratamento para cada fator, teremos apenas 1 grau de liberdade para a soma dos quadrados da interação – não importa quanta replicação tenhamos usado (*ver* Nota de rodapé 10, Capítulo 7).

O esforço amostral apropriado para um modelo de efeitos fixos *versus* um de efeitos aleatórios deveria fazer algum sentido intuitivo. Se o domínio de inferência é restrito aos tratamentos estabelecidos, então uma replicação extra aumentará nosso poder no modelo de efeitos fixos. Mas se os níveis de tratamento são apenas subconjuntos aleatórios de muitos níveis de tratamento possíveis, devemos tentar aumentar a cobertura dos diferentes níveis para o modelo de efeitos aleatórios. Se os níveis de tratamento representam uma variável contínua que foi convertida

Quadrado médio esperado	Razão-F	Valor de *P*
$\sigma^2 + n\sigma^2_{AB} + nb\sigma^2_A$	$\frac{QM_A}{QM_{AB}}$	Cauda da distribuição F com $(a-1)$, $(a-1)(b-1)$ graus de liberdade
$\sigma^2 + n\sigma^2_{AB} + na\sigma^2_B$	$\frac{QM_B}{QM_{AB}}$	Cauda da distribuição F com $(b-1)$, $(a-1)(b-1)$ graus de liberdade
$\sigma^2 + n\sigma^2_{AB}$	$\frac{QM_{AB}}{QM_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)(b-1)$, $ab(n-1)$ graus de liberdade
σ^2		
σ^2_Y		

em um fator discreto para a ANOVA, a melhor estratégia pode ser abandonar por completo os delineamentos de ANOVA e tratar o problema como um delineamento experimental de regressão ou de superfície de resposta, conforme discutimos no Capítulo 7.

Nem sempre é fácil decidir se um fator deve ser analisado como fixo ou aleatório. Se trata de um conjunto de locais ou de tempos escolhidos de maneira aleatória, quase sempre deve ser tratado como um fator aleatório, que é como os blocos são analisados em um delineamento de blocos aleatorizados ou em um delineamento de medidas repetidas. Se ele representa um conjunto de categorias bem-definidas que são limitadas em número e não representa uma variação contínua, como espécies ou sexos, esse fator provavelmente deve ser tratado como do tipo fixo.

Uma abordagem é considerar a razão x/X, onde x é o número de tratamentos no estudo e X é o número de níveis de tratamento possíveis. Se esta razão der próximo de 0, você provavelmente tem um fator aleatório; se der próximo de 1,0, você provavelmente tem um fator fixo.

Muitas variáveis naturalmente contínuas, como a concentração de nutrientes ou densidade populacional, deveriam ser analisadas como fatores aleatórios se você estabelecer níveis discretos para o seu delineamento de ANOVA. Entretanto, se há

TABELA 10.10 Tabela de ANOVA para um modelo misto de dois fatores

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)
Fator A	$a-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_i-\overline{Y})^2$	$\frac{SQ_A}{(a-1)}$
Fator B	$b-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_j-\overline{Y})^2$	$\frac{SQ_B}{(b-1)}$
Interação ($A\times B$)	$(a-1)(b-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{ij}-\overline{Y}_i-\overline{Y}_j+\overline{Y})^2$	$\frac{SQ_{AB}}{(a-1)(b-1)}$
Dentro de grupos (resíduos)	$ab(n-1)$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{ij})^2$	$\frac{SQ_{dentro\ de\ grupos}}{ab(n-1)}$
Total	$abn-1$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y})^2$	$\frac{SQ_{total}}{(abn-1)}$

O Fator A é fixo e o B é aleatório. Há de $i = 1$ até a níveis do Fator A, de $j = 1$ até b níveis do Fator B e n réplicas de cada combinação de tratamentos única ij. Note que a razão-F para o Fator A utiliza o quadrado médio da interação no denominador, mas a razão-F para o Fator B e para o efeito de interação ($A \times B$) utilizam o quadrado médio dos resíduos no denominador. Os graus de liberdade para esses dois efeitos também são diferentes.

uma significância especial para os níveis de tratamento em particular que você utilizou, pode ser melhor utilizar um delineamento fixo. Por exemplo, em um estudo de densidade, você pode estabelecer apenas dois tratamentos: densidade ambiental e densidade zero. Embora haja muitos outros níveis possíveis, o primeiro pode ser visto como condições em equilíbrio, enquanto a ausência da espécie representa uma condição de pré-invasão. Um delineamento fixo poderia ser apropriado neste caso.

Você deve deixar claro na seção dos métodos do seu artigo ou relatório quais fatores foram fixos e/ou aleatórios e utilizar o delineamento apropriado para a sua ANOVA. Não há desculpa para confiar nas análises de fatores fixos padrão, especialmente quando a maioria dos programas estatísticos permite que você especifique quais termos de erro serão utilizados. As Tabelas 10.9 e 10.10 ilustram o quadrado médio esperado apropriado apenas para o modelo de ANOVA de dois fatores. Para um delineamento mais complexo, os quadrados médios esperados para efeitos mistos e aleatórios podem ser encontrados em outros textos sobre ANOVA (p. ex., Winer et al., 1991). Entretanto, em alguns delineamentos complexos pode não ser possível construir razões-F válidas para modelos mistos. Essa é outra razão para usar delineamentos de ANOVA simples e evitar delineamentos que utilizam muitos fatores com aninhamento ou medidas repetidas em suas análises.

Quadrado médio esperado	Razão-F	Valor de *P*
$\sigma^2 + n\sigma^2_{AB} + nb\sigma^2_A$	$\frac{QM_A}{QM_{AB}}$	Cauda da distribuição F com $(a-1), (a-1)(b-1)$ graus de liberdade
$\sigma^2 + na\sigma^2_B$	$\frac{QM_B}{QM_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(b-1), ab(n-1)$ graus de liberdade
$\sigma^2 + n\sigma^2_{AB}$	$\frac{QM_{AB}}{QM_{dentro\ de\ grupos}}$	Cauda da distribuição F com $(a-1)(b-1), ab(n-1)$ graus de liberdade
σ^2		
σ^2_Y		

PARTIÇÃO DE VARIÂNCIA NA ANOVA

Muitos usos da ANOVA em ecologia e ciências ambientais têm o objetivo de testar hipóteses sobre os efeitos principais e termos de interação. Entretanto, como destacamos no Capítulo 4, testar hipóteses é apenas o primeiro passo ao analisar dados. Também estamos interessados em estimar parâmetros de modelos ou em determinar qual variação nos dados pode ser atribuída a uma fonte em particular (Figura 4.6). A noção de partição de variância é familiar da análise de regressão (Capítulo 9), na qual o valor do coeficiente de determinação (r^2) pode ser mais importante que o fato de a hipótese nula ser rejeitada ou não.

Podemos fazer a partição da variância na ANOVA de maneira semelhante, mas o procedimento não é matematicamente tão simples quanto na regressão. No passado, os pesquisadores tentaram inferir, incorretamente, a importância relativa de um fator em particular a partir do tamanho do quadrado médio ou do valor de *P* calculado. Ambas as abordagens estão erradas. O quadrado médio quase sempre estima mais de um componente de variância, portanto ele não pode ser utilizado sozinho. Além disso, o valor de *P* será sensível ao tamanho amostral utilizado no estudo.

A abordagem correta é começar com o quadrado médio esperado da sua tabela de ANOVA em particular e depois transformá-lo algebricamente para isolar o componente de variância de interesse. Por exemplo, em uma ANOVA com um fator fixo, sabemos que o quadrado médio dos resíduos estima a variação dentro de grupos:

$$\sigma_e^2 = QM_{dentro\ de\ grupos} \qquad \textbf{(10.19)}$$

O quadrado médio entre grupos estima o efeito do tratamento mais o resíduo:

$$\sigma_e^2 + n\sigma_A^2 = QM_{entre\ grupos} \qquad \textbf{(10.20)}$$

Se rearranjarmos a Equação 10.20, podemos isolar o efeito do tratamento:

$$\sigma_A^2 = \frac{QM_{entre\ grupos} - QM_{dentro\ de\ grupos}}{n} \qquad \textbf{(10.21)}$$

Um refinamento adicional é necessário. Como essa é uma ANOVA de fator fixo, há um número finito de níveis de tratamento do Fator *A*. Portanto, a estimativa da variância devido ao efeito do tratamento (que, a rigor, não é uma "variância" verdadeira) precisa ser ajustada pelo fator $(a - 1)/a$. Esse ajuste leva em conta a diferença nos graus de liberdade que temos para uma medida de variância verdadeira (a) e uma estimativa de variância a partir de um amostragem finita

$(a - 1)$. Note que conforme o número de grupos de tratamento aumenta, o fator de correção $[(a - 1)/a]$ se aproxima de 1,0. Isso tem sentido intuitivo, porque um modelo de efeitos fixos com um grande número de níveis de tratamento efetivamente é um modelo de efeitos aleatórios.

Com esse pouco de contabilidade para os componentes de fatores fixos, temos:

$$\sigma_A^2 = \frac{(QM_{entre\ grupos} - QM_{dentro\ de\ grupos})(a-1)}{na} \qquad (10.22)$$

Uma vez que os componentes de variância tenham sido isolados dessa maneira, suas contribuições para o total podem ser calculadas como a **proporção de variância explicada (PVE)**:

$$PVE_e = \frac{\sigma_e^2}{\sigma_A^2 + \sigma_e^2} \qquad (10.23)$$

e

$$\text{PVE}_A = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2} \qquad (10.24)$$

Algebricamente, essa partição equivale ao cálculo do r^2 em uma regressão linear (Capítulo 9).

Como exemplo, usando os dados da Tabela 10.1, o $QM_{entre\ grupos} = 11{,}085$, o $QM_{dentro\ de\ grupos} = 2{,}167$, $a = 3$ grupos, $n = 4$ réplicas por grupo. A partir das Equações 10.21 e 10.22, a estimativa da variância para o efeito do tratamento é 1,486 e a estimativa da variância para o resíduo é 2,167. Pela Equação 10.23, a PVE_A = 41%: (1,486)/(1,486 + 2,167) = 0,406.

Para ANOVAs de dois fatores, seguimos o mesmo princípio de isolar algebricamente o termo de interesse, embora os cálculos sejam um pouco mais complexos. Como você pode imaginar, os cálculos na ANOVA de dois fatores dependem de se estamos utilizando modelos de efeitos fixos, aleatórios ou mistos (Tabela 10.11).

Mesmo com os cálculos corretos, a técnica ainda tem algumas limitações importantes. Talvez o problema conceitual mais sério é que a partição se aplica apenas aos fatores que foram realmente medidos e incluídos no modelo. Qualquer variação em fatores não medidos, independente de sua importância, é inserida no quadrado médio do resíduo ou nos efeitos do tratamento medido (se houver interações entre os fatores medidos e os não medidos).

Mesmo em uma ANOVA de um fator, a quantidade de variação atribuível ao efeito do tratamento *versus* ao resíduo ou erro dependerá não apenas do número

TABELA 10.11 Componentes de variância em modelos de ANOVA de dois fatores com fatores fixos, aleatórios ou mistos

Componente de variância	Modelo de efeitos fixos (*A* fixo, *B* fixo)	Modelo de efeitos aleatórios (A aleatório, *B* aleatório)
A	$\frac{(QM_A - QM_{resíduo})(a-1)}{abn}$	$\frac{(QM_A - QM_{A\times B})}{bn}$
B	$\frac{(QM_B - QM_{resíduo})(b-1)}{abn}$	$\frac{(QM_B - QM_{A\times B})}{an}$
$A \times B$	$\frac{(QM_{A\times B} - QM_{resíduo})(a-1)(b-1)}{abn}$	$\frac{(QM_{A\times B} - QM_{resíduo})}{n}$
Resíduo	$QM_{resíduo}$	$QM_{resíduo}$

de grupos de tratamentos utilizados, mas também dos níveis em particular que são escolhidos (Underwood & Petraitis, 1993; Petraitis, 1998). Similarmente, em um estudo de regressão com uma variável preditora contínua, a quantidade de variação contabilizada dependerá da amplitude dos valores de X escolhidos (Figuras 7.2 e 7.3).

Esta limitação indica que é difícil, mas não impossível, comparar a intensidade de forças relativas em diferentes estudos.[7] Entretanto, as mesmas ressalvas se aplicam na interpretação dos valores de P calculados na ANOVA. Quando calculada corretamente e interpretada de forma conservativa, a partição de variância é um complemento útil para o teste de hipóteses padrão da ANOVA.

APÓS A ANOVA: PLOTANDO E ENTENDENDO OS TERMOS DE INTERAÇÃO

Neste ponto, você já especificou sua questão científica (Capítulo 4), delineou e executou seu experimento (Capítulos 6 e 7), coletou e gerenciou seus dados (Ca-

[7] Uma abordagem completamente diferente para avaliar estudos de efeitos entre estudos é a meta-análise (Arnqvist & Wooster, 1995). Na meta-análise, cada estudo individual é tratado como uma réplica independente para o teste de uma hipótese mais ampla (p. ex., cascatas tróficas são mais intensas em cadeias alimentares aquáticas *versus* terrestres; Shurin et al., 2002). A partir de estudos publicados, a meta-análise extrai um tamanho do efeito, em geral comparando as médias do controle e dos grupos de tratamentos, padronizados pela variância estimada. A meta-análise tem se tornado popular recentemente, porém controversa, na ecologia e evolução. Quantificar o tamanho do efeito apropriadamente (Osenberg et al., 1999) e controlar os vieses das publicações (Jennions & Møller, 2002) são dois desafios atuais na meta-análise. *Ver* Gurevitch et al. (2001) para uma introdução aos métodos.

Modelo de efeitos mistos (A fixo, *B* aleatório)

$$\frac{(QM_A - QM_{A\times B})(a-1)}{bna}$$

$$\frac{(QM_B - QM_{A\times B})}{an}$$

$$\frac{(QM_{A\times B} - QM_{resíduo})}{n}$$

$$QM_{resíduo}$$

Aqui, o a é o número de níveis de tratamento no Fator A; b é o número de níveis de tratamento no Fator B; n é o número de réplicas por tratamento (em um delineamento balanceado); QM indica um valor de quadrado médio em particular, que é usado a partir da tabela de ANOVA. Como o quadrado médio esperado é diferente em modelos de ANOVA com fatores fixos, aleatórios e mistos, a estimativa dos componentes de variância difere para esses modelos. Uma vez que os componentes de variância são estimados, podem ser usados para medir a porcentagem de variação a ser atribuída para cada fator na análise.

pítulo 8) e os analisou com ANOVA, usando métodos frequentistas, de Monte Carlo ou bayesianos (Capítulo 5).

O próximo passo crítico é plotar os resultados das análises. É impossível interpretar estatística ou biologicamente os resultados de uma ANOVA sem uma referência cuidadosa aos gráficos dos dados, e você nunca deve publicar apenas uma tabela de ANOVA sem publicar, também, alguma medida do tamanho dos efeitos (p. ex., médias dos grupos e das variâncias), ou em tabelas ou em gráficos. Já discutimos a análise de dados exploratória (EDA) como forma de checar a presença de dados discrepantes ou erros (Capítulo 8) e de examinar os padrões gerais nos seus dados. Aqui, enfatizamos gráficos com qualidade de publicação que destacam os padrões das médias e variâncias, e que permitem que você interprete os resultados dos testes estatísticos usando a ANOVA.

Plotando resultados de ANOVAs de um fator

Vamos considerar o delineamento de um fator mais simples para um estudo experimental genérico: três grupos tratamentos, constituídos por um conjunto de parcelas sem manipulação (S) que não são alteradas de nenhuma forma, exceto pelo censo ou pelas medições que são realizadas; um conjunto de parcelas controle (C) que incorpora possíveis efeitos da manipulação, mas que na realidade não implementa o tratamento de interesse; e as parcelas de tratamento (T), que implementam o tratamento de interesse e (por necessidade) também incluem os efeitos da manipulação.

Podemos plotar esses dados em simples gráfico de barras. O eixo-y representa a variável resposta, plotada em suas unidades de medida. No eixo-x, há um ró-

tulo para cada grupo de tratamento plotado (3, nesse caso). Para cada um, plote uma barra simples (ou um gráfico de barras, *box-and-whisker*; *ver* os Capítulos 3 e 8). A altura da barra representa a média do grupo de tratamento. Adicione uma linha vertical acima da barra para indicar o desvio-padrão ao redor da média. Idealmente, a média amostral e o desvio-padrão amostral deveriam basear-se em, no mínimo, 10 observações em cada grupo (*ver* Capítulo 6, "A Regra dos 10"). Para delineamentos mais complicados, você pode utilizar sombreamento para indicar tratamentos que representem subgrupos importantes nos seus dados. Esses subgrupos podem ser comparados com contrastes *a priori*, discutidos a seguir, neste capítulo.

Com esse gráfico em mãos, agora podemos interpretar nossos resultados da ANOVA. Se o teste da razão-F para diferenças entre grupos não for significativo, quer dizer que não há evidência de que a população de médias seja diferente do que poderia ser esperado ao acaso, devido ao erro amostral aleatório (Figura 10.2A). A ressalva é que o nosso experimento ou delineamento seguiu os pressupostos da ANOVA e que nosso tamanho amostral é suficientemente grande para ter um poder razoável para detectar um efeito (*ver* Capítulo 4). Note que se a variância dentro de grupos é grande e o tamanho amostral pequeno, as médias amostrais podem ser ligeiramente diferentes umas das outras, mesmo que a hipótese nula não tenha sido rejeitada. Por isso é tão importante plotar os desvios-padrão junto com as médias amostrais e considerar, com cuidado, o tamanho amostral do experimento.

E se o teste da razão-F para diferenças entre os grupos é significativo? Existem três padrões gerais nas médias que levam a três diferentes interpretações. Primeiro, suponha que a média das parcelas do tratamento seja elevada (ou reduzida) se comparada às médias das parcelas-controle e das sem manipulação (T > C = S, Figura 10.2B). Esse resultado sugere que o tratamento tem um efeito significativo na variável resposta e que os resultados não representam um artefato do delineamento. Note que o efeito do tratamento poderia ter sido inferido erroneamente caso tivéssemos (por equívoco) executado o experimento sem os controles e comparado apenas os grupos de tratamento e os sem manipulação. Como alternativa, suponha que os grupos de tratamento e os controles tenham médias similares, mas elas são mais elevadas que a das parcelas sem manipulação (T = C > S; Figura 10.2C). Esse padrão sugere que o efeito do tratamento não é importante e que a diferença entre as médias reflete o efeito da manipulação ou outro artefato da manipulação. Por fim, suponha que as três médias diferem entre si, com a do grupo de tratamento sendo a maior e a do grupo sem manipulação a menor (T > C > S; Figura 10.2D). Neste caso, há evidência de um efeito da manipulação, pois C > S. Contudo, o fato de T > C significa que o efeito do tratamento também é real e não representa apenas artefatos da manipulação. Uma vez que tenha estabelecido o padrão em seus dados usando a ANOVA e os resultados gráficos, o próximo passo é determinar se o padrão tende a suportar ou refutar as hipóteses científicas em avaliação (Capítulo 4). Sempre tenha em mente a diferença entre as significâncias estatística e biológica!

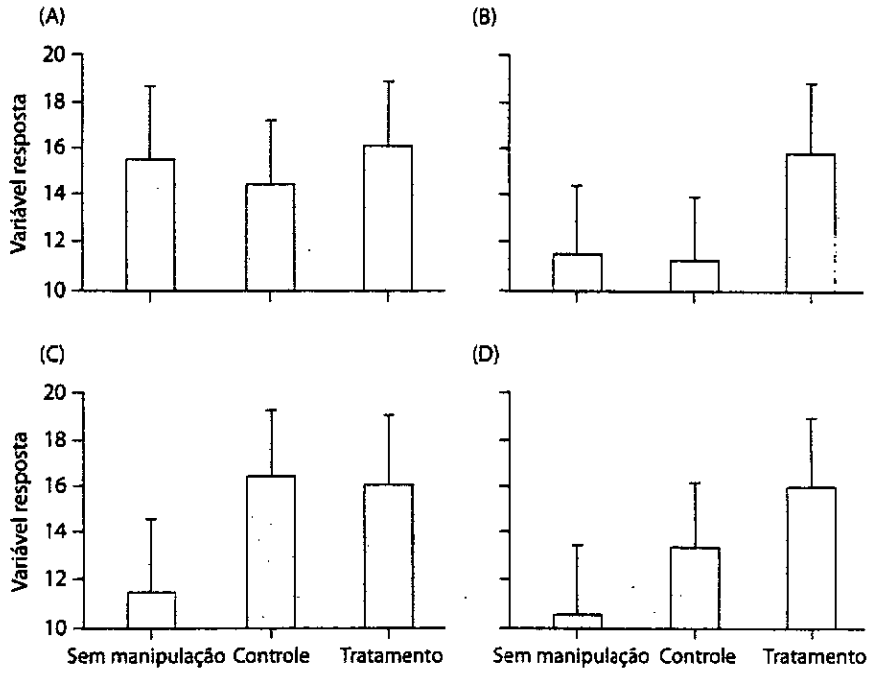


Figura 10.2 Possíveis resultados de um experimento hipotético com três tratamentos e uma configuração de ANOVA de um fator. Parcelas sem manipulação não são alteradas em nenhuma forma, exceto para os efeitos que ocorrem durante a amostragem. As parcelas-controle não recebem o tratamento, mas podem receber um tratamento falso para imitar os efeitos da manipulação. O grupo de tratamento possui a manipulação de interesse, mas potencialmente também inclui os efeitos de manipulação. A altura de cada barra representa a resposta média para cada grupo e as linhas verticais indicam um desvio-padrão ao redor da média. (A) O teste de ANOVA para os efeitos do tratamento não é significativo e as diferenças entre grupos na resposta média pode ser atribuída apenas ao erro. (B-D) O teste de ANOVA é sempre significativo, mas o padrão – e,portanto, a interpretação – difere em cada caso. (B) A média do tratamento é elevada se comparada às médias das réplicas controle e sem manipulação, indicando um efeito real do tratamento. (C) Tanto as médias dos grupos controle quanto as do tratamento são elevadas em relação às parcelas sem manipulação, indicando um efeito da manipulação, mas não distinto do tratamento. (D) Há evidência de um efeito da manipulação, pois a média do grupo controle excede à das parcelas sem manipulação, embora o tratamento desejado (biológico) não tenha sido aplicado. No entanto, para além desse efeito de manipulação, parece haver o do tratamento, pois a média do grupo de tratamento é elevada em relação à do grupo controle. Contrastes *a priori* ou comparações *a posteriori* podem ser utilizadas para apontar quais grupos são distintamente diferentes uns dos outros. Nosso ponto, aqui, é que os resultados da ANOVA podem ser idênticos, mas a interpretação depende do padrão das médias dos grupos em particular. Os resultados da ANOVA não podem ser interpretados corretamente sem um exame cuidadoso do padrão das médias e variâncias dos grupos de tratamento.

Plotando resultados de ANOVAs de dois fatores

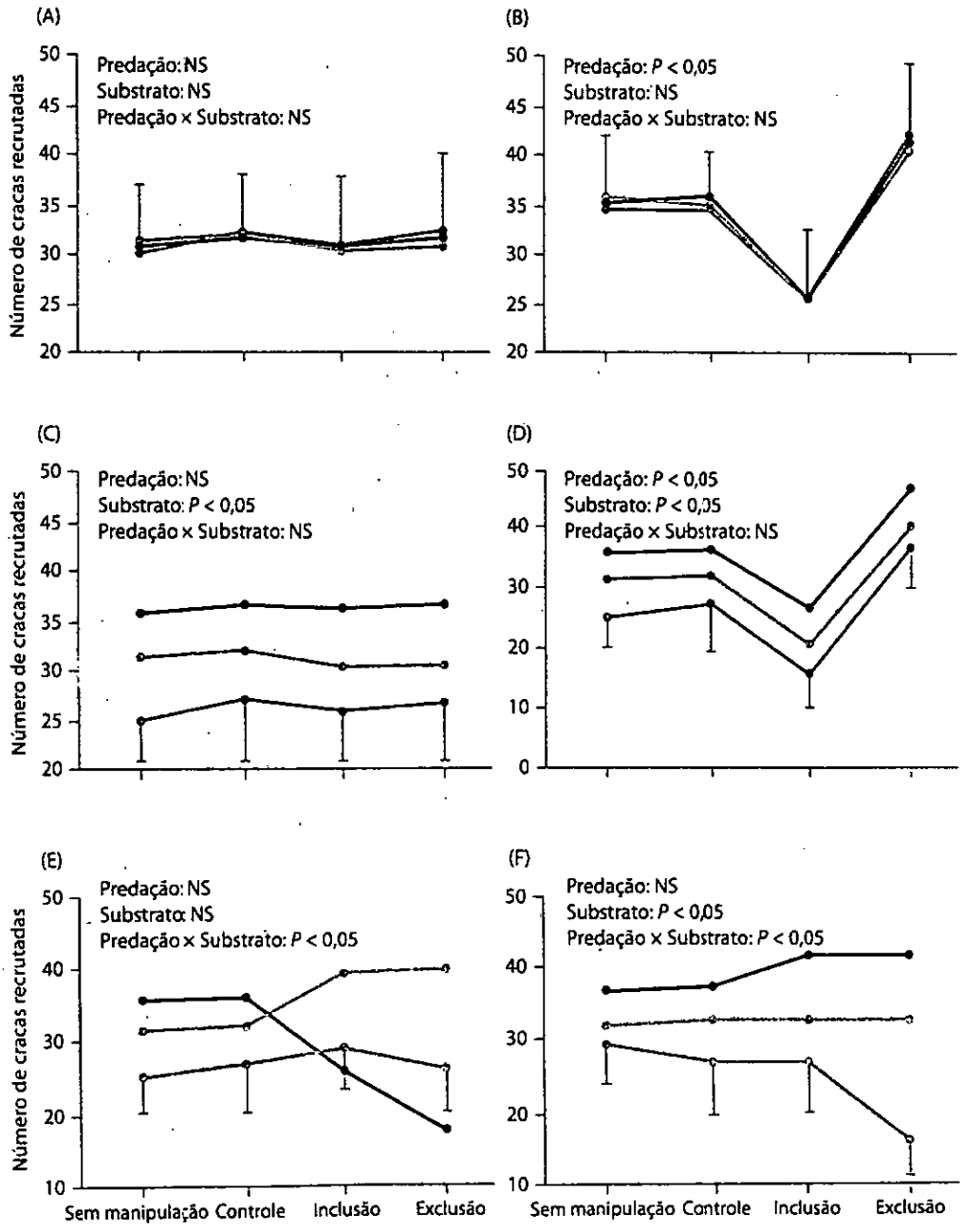
Para plotar os resultados de ANOVAs de dois fatores, gráficos de barras simples também podem ser utilizados, mas eles são mais difíceis de interpretar grafica-

mente e não mostram muito bem os efeitos principais e os de interação. Sugerimos o seguinte protocolo para plotar as médias de uma ANOVA de dois fatores:

1. Estabeleça um gráfico no qual o eixo-*y* seja a variável resposta contínua e o eixo-*x*, os diferentes níveis do primeiro fator do experimento.
2. Use um símbolo ou uma cor diferente para cada nível do tratamento ao representar o segundo fator do experimento. Cada símbolo, colocado sobre o rótulo da categoria apropriada no eixo-*x*, representa a média de uma combinação de tratamentos em particular. Haverá um total de $a \times b$ símbolos, onde a é o número de tratamentos do primeiro fator e b é o de tratamentos do segundo fator.
3. Alinhe cada símbolo acima do rótulo do fator apropriado para estabelecer a combinação de tratamentos representada.
4. Use uma linha (colorida) para conectar os símbolos através dos níveis do primeiro fator.
5. Para plotar os desvios-padrão, use linhas verticais apontando para cima (para os níveis de tratamento superiores) e linhas verticais apontando para baixo (níveis de tratamento inferiores). Se a figura se tornar muito desorganizada, as barras de erro podem ser ilustradas apenas para algumas das séries de dados.

Figura 10.3 Possíveis resultados para um experimento hipotético testando efeitos dos tratamentos de substrato e de predação sobre o recrutamento de cracas. Cada símbolo representa a média para diferentes combinações de tratamentos. Os tratamentos de predação são indicados pelo rótulo no eixo-*x* e os dos substratos são indicados por cores diferentes (verde = cimento; cinza = ardósia; preto = granito). As barras verticais indicam o erro-padrão ou desvio-padrão (o que precisa ser apontado na legenda da figura). Cada painel representa um padrão associado a um resultado estatístico em particular. (A) Nem os efeitos dos tratamentos nem o da interação são significativos e as médias de todas as combinações de tratamentos são estatisticamente indistinguíveis. (B) O efeito da predação é significativo, mas o do substrato e o da interação não são. Neste gráfico, as médias dos tratamentos são maiores para a exclusão do predador e menores para a sua inclusão, com um padrão similar para todos os substratos. (C) O efeito do substrato é significativo, mas os do predador e da interação não. Não há diferenças nas médias do tratamento de predação, mas o recrutamento é sempre maior nos substratos de granito e menor nos de cimento, independente do tratamento de predação. (D) Tanto o efeito de predação quanto do substrato são significativos, mas o efeito de interação não é. As médias dependem tanto do tratamento de substrato quanto do de predação, porém o efeito é estritamente aditivo e os perfis das médias são paralelos ao longo dos substratos e dos tratamentos. (E) Efeito de interação significativo. As médias dos tratamentos diferem de maneira significativa, mas já não há um efeito aditivo simples da predação ou do substrato. O *ranking* das médias dos substratos depende do tratamento de predação, e o *ranking* das médias de predação, do substrato. Os efeitos principais podem não ser significativos nesse caso, porque as médias dos tratamentos de substrato ou de predação não necessariamente diferem de maneira significativa. (F) O efeito da interação é marcante, o que significa que o efeito do tratamento de predação depende do tratamento de substrato e vice-versa. Apesar dessa interação, ainda é possível falar sobre os efeitos gerais do substrato sobre o recrutamento. Independente do tratamento de predação, o recrutamento sempre é maior em substratos de granito e menor nos de cimento. O efeito da interação é estatisticamente significativo, mas os perfis das médias, na verdade, não se cruzam.

A Figura 10.3 mostra exatamente esse tipo de gráfico para o experimento hipotético de dois fatores das cracas que descrevemos. O eixo-*x* dá os quatro níveis do tratamento de predação (sem manipulação, controle, exclusão do predador e inclusão do predador). Três cores (verde, cinza e preto) são usadas para indicar os três níveis do



tratamento de substrato (cimento, ardósia e granito). Para cada tipo de substrato, as médias dos quatro níveis do tratamento de predação são conectadas por uma linha.

Mais uma vez, devemos considerar os diferentes resultados da ANOVA e o que os gráficos associados poderiam parecer. Neste caso, existem várias possibilidades, pois agora temos dois testes de hipóteses para os efeitos principais da predação e do substrato, e um terceiro teste de hipóteses para a interação entre os dois.

NENHUM EFEITO SIGNIFICATIVO Como antes, o cenário mais simples é aquele no qual nenhum dos dois efeitos principais nem o termo de interação são estatisticamente significativos. Se os tamanhos amostrais forem razoavelmente grandes, este também será o caso mais confuso para plotar, pois todas as médias dos grupos de tratamento serão muito similares entre si e os pontos das médias podem ficar bastante sobrepostos (Figura 10.3A).

UM EFEITO PRINCIPAL SIGNIFICATIVO A seguir, suponha que o efeito principal de predação é significativo, mas o do substratos e da interação não. Portanto, as médias de cada tratamento de predação, calculadas sem a dos três tratamentos de substratos, são significativamente diferentes umas das outras. Em contrapartida, as médias dos tipos de substratos, calculadas sem a média dos diferentes tratamentos de predação não são substancialmente diferentes umas das outras. O gráfico mostrará aglomerações de médias diferentes em cada nível do tratamento de predação, mas as de cada tipo de substrato serão praticamente idênticas conforme o nível de predação (Figura 10.3B).

Do contrário, suponha que o efeito do substrato seja significativo, mas que o efeito da predação não. Agora, as médias de cada um dos três tratamentos de substrato são significativamente diferentes, calculadas pela média dos quatro tratamentos de predação. Contudo, as dos tratamentos de predação não diferem. As três linhas conectadas para os tipos de substratos serão bem separadas umas das outras, mas as suas inclinações serão basicamente planas, porque não há efeito dos quatro tratamentos (Figura 10.3C).

DOIS EFEITOS PRINCIPAIS SIGNIFICATIVOS A próxima possibilidade é um efeito significativo da predação e do substrato, mas não do termo de interação. Neste caso, o perfil das respostas médias novamente é diferente para cada tipo de substrato, mas ele não é mais plano ao longo dos tratamentos de predação. Uma característica-chave desse gráfico é que as linhas que conectam os diferentes grupos de tratamentos são paralelas umas às outras. Quando os perfis são paralelos, os efeitos dos dois fatores são estritamente aditivos: uma combinação dos tratamentos em particular pode ser predita sabendo os efeitos médios de cada um dos dois tratamentos individuais. A aditividade dos efeitos do tratamento e um perfil paralelo dos tratamentos são diagnósticos para uma ANOVA na qual ambos os efeitos principais são significativos e o termo de interação não é (Figura 10.3D).

EFEITO DA INTERAÇÃO SIGNIFICATIVO A última possibilidade que consideraremos é quando o termo de interação é significativo, mas nenhum dos dois efeitos principais o são. Sempre que houver um termo de interação significativo, as linhas dos gráficos de perfil já não serão (estatisticamente) paralelas (Figura 10.3F) e podem até mesmo se cruzarem caso as interações sejam fortes (Figura 10.3E).

Entendendo os termos de interação

Com um termo de interação significativo, as médias dos grupos de tratamento são bastante diferentes umas das outras, mas não podemos mais descrever um efeito aditivo simples para nenhum dos dois fatores no delineamento. Em vez disso, o efeito do primeiro fator (p. ex., predação) depende do nível do segundo (p. ex., tipo de substrato). Deste modo, as diferenças entre os tipos de substrato dependem de qual tratamento de predação está sendo considerado. Nas parcelas controle e nas sem manipulação, o recrutamento foi maior em substratos de granito; em parcelas com inclusão e exclusão de predadores, o recrutamento foi maior em ardósia. Equivalentemente, podemos dizer que o efeito do tratamento de predação depende do substrato. Para substratos de granito, a abundância é maior nos controles; em substratos de ardósia, é maior em tratamentos de inclusão e exclusão de predadores.

A representação gráfica do termo de interação, como parcelas não paralelas em relação às médias dos tratamentos, também possui uma interpretação algébrica. A partir da Tabela 10.6, a soma dos quadrados para o termo de interação em uma ANOVA de dois fatores é:

$$SQ_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\overline{Y}_{ij} - \overline{Y}_i - \overline{Y}_j + \overline{Y})^2 \qquad (10.25)$$

Equivalentemente, podemos adicionar e subtrair um termo $\overline{Y}$, dando:

$$SQ_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left[(\overline{Y}_{ij} - \overline{Y}) - (\overline{Y}_i - \overline{Y}) - (\overline{Y}_j - \overline{Y})\right]^2 \qquad (10.26)$$

O primeiro termo $\overline{Y}_{ij} - \overline{Y}$, nessa expressão expandida, representa o desvio da média de cada grupo de tratamento para a média geral. O segundo termo $\overline{Y}_i - \overline{Y}$ representa o desvio do efeito aditivo do Fator *A*. O terceiro termo $\overline{Y}_j - \overline{Y}$ representa o efeito aditivo do Fator *B*. Se os efeitos aditivos dos Fatores *A* e *B* juntos descrevem todos os desvios das médias dos tratamentos para a média geral, o efeito da interação é zero. Assim, o termo de interação mede a extensão na qual as médias dos tratamentos diferem dos efeitos estritamente aditivos dos dois fatores principais. Se não houver termo de interação, saber o tipo de substrato e o tratamento de predação nos permite predizer com perfeição a resposta quando esses dois fatores forem combinados. Mas se houver uma interação forte, já não poderíamos

mais predizer os efeitos combinados, mesmo se entendêssemos o comportamento dos fatores separadamente.[8]

É evidente o motivo de o termo de interação ser significativo na Figura 10.3E, mas por que os efeitos principais não deveriam ser importantes neste caso? A razão é: se tirar a média das médias ao longo de cada tratamento de predação ou cada tipo de substrato, elas seriam aproximadamente iguais e não haveriam diferenças consistentes para cada fator considerado de forma isolada. Por esta razão, as vezes alega-se que nada pode ser dito sobre os efeitos principais quando os termos de interação são significativos. Esta afirmação é verdadeira somente para interações muito fortes, nas quais o perfil das curvas se cruzam. Entretanto, em muitos casos pode haver tendências gerais para um único fator, mesmo quando o termo de interação é significativo.

Por exemplo, considere a Figura 10.3F, que certamente poderia gerar um termo de interação estatisticamente significativo em uma ANOVA de dois fatores. Essa interação nos diz que as diferenças entre as médias para os tipos de substratos dependem do tratamento de predação. Entretanto, neste caso, a interação surge principalmente porque o tratamento com substrato de cimento × exclusão do predador possui uma média muito pequena em relação aos outros tratamentos. Em todos os tratamentos de predação, o substrato de granito (linha preta) possui o maior recrutamento e o substrato de cimento (linha verde), o menor. No tratamento controle, as diferenças entre as médias dos substratos são relativamente pequenas; enquanto no tratamento de exclusão do predador, as diferenças entre médias dos substratos são relativamente grandes. Novamente, enfatizamos que os resultados da ANOVA não podem ser interpretados de maneira adequada sem referência aos padrões das médias e às variâncias dos dados.

Por fim, note que as transformações dos dados algumas vezes podem eliminar os termos de interação significativos. Em particular, as relações que são multiplicativas em uma escala linear são aditivas na escala logarítmica (*ver* discussão nos Capítulos 3 e 8) e a transformação logarítmica pode, com frequência, eliminar os termos de interação significativos. Certos delineamentos de ANOVA não incluem termos de interação e temos de assumir aditividade estrita nesses modelos.

Plotando resultados de ANCOVAs

Concluiremos esta seção considerando os possíveis resultados de uma ANCOVA. O gráfico da ANCOVA deve utilizar a covariável contínua plotada no eixo-*x* e a

[8] Um exemplo mórbido de interação estatística é o efeito do álcool e dos sedativos na pressão sanguínea humana. Suponha que o álcool diminua a pressão sanguínea em 20 pontos, e os sedativos, em 15 pontos: em um mundo aditivo simples, a combinação de álcool e sedativos deveria diminuir a pressão sanguínea em 35 pontos. Em vez disso, a interação entre álcool e sedativos pode diminuir a pressão sanguínea em 50 pontos ou mais e isso frequentemente é letal. Este resultado não poderia ser predito apenas entendendo seus efeitos em separado. As interações são um problema sério, tanto na medicina quanto nas ciências ambientais. Os estudos experimentais simples podem quantificar os efeitos de um único estressante ambiental, como as elevações de CO_2 ou o aumento da temperatura, mas pode haver fortes interações entre esses fatores que causem resultados inesperados.

variável *Y*, no eixo-*y*. Cada ponto representa uma réplica independente e símbolos e cores diferentes devem ser utilizados para indicar réplicas de diferentes tratamentos. Inicialmente, ajuste uma linha de regressão linear para cada grupo de tratamento em um gráfico geral de resultados. Apresentaremos um único exemplo onde há uma covariável e três níveis de tratamento sob comparação. Eis as possibilidades:

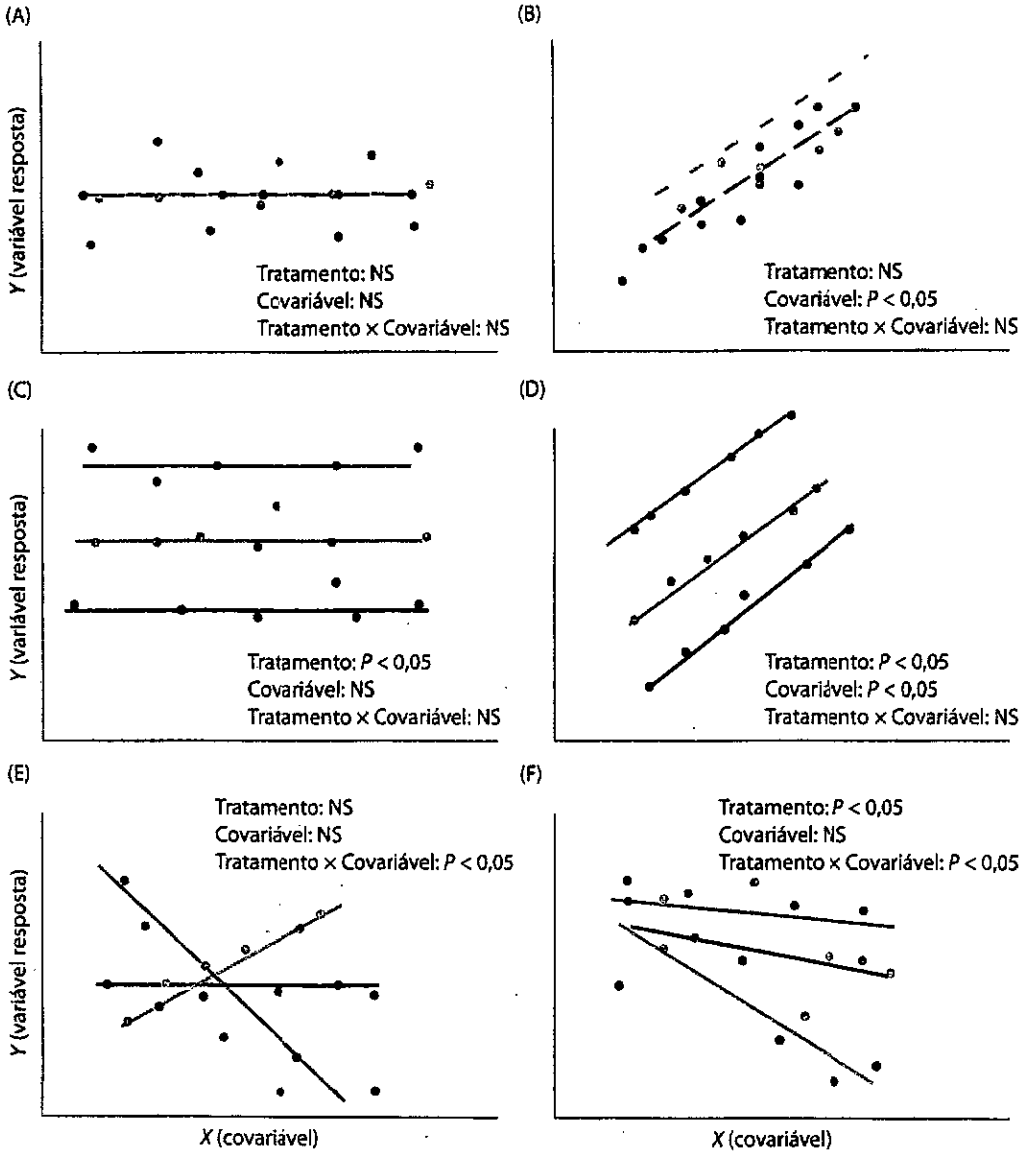
COVARIÁVEL, TRATAMENTO E INTERAÇÃO NÃO SIGNIFICATIVOS Neste caso, as três linhas de regressão não diferem umas das outras e a sua inclinação não é diferente de zero. O intercepto ajustado para cada regressão estima efetivamente a média dos valores *Y* (Figura 10.4A).

COVARIÁVEL SIGNIFICATIVA, TRATAMENTO E INTERAÇÃO NÃO SIGNIFICATIVOS Neste caso, as três linhas de regressão não diferem umas das outras, mas agora a inclinação da linha de regressão universal é significativamente diferente de zero. Os resultados sugerem que a covariável explica a variação nos dados, mas não há diferenças entre os tratamentos após o efeito da covariável ser (estatisticamente) removido (Figura 10.4B).

TRATAMENTO SIGNIFICATIVO, COVARIÁVEL E TERMO DE INTERAÇÃO NÃO SIGNIFICATIVOS Neste caso, as linhas de regressão novamente possuem inclinação equivalente a zero, mas agora o intercepto dos três níveis pode ser diferente, indicando um efeito significativo do tratamento. Como a covariável não explica muito sobre a variação nos dados, o resultado poderia ser qualitativamente o mesmo caso uma ANOVA de um fator fosse usada e as medidas da covariável, ignoradas (Figura 10.4C).

TRATAMENTO E COVARIÁVEL SIGNIFICATIVOS E TERMO DE INTERAÇÃO NÃO SIGNIFICATIVO Neste caso, as linhas de regressão são equivalentes, possuem inclinações diferentes de zero, e os interceptos são significativamente diferentes. A covariável explica parte da variação nos dados, mas ainda há variação do resíduo que pode ser atribuída ao efeito do tratamento. Este resultado seria equivalente a ajustar uma única linha de regressão ao conjunto de dados inteiro e depois usar os resíduos para testar os efeitos dos tratamentos com uma ANOVA de um fator. Quando esse resultado é obtido, é apropriado ajustar a inclinação da regressão universal e utilizar a Equação 10.18 para estimar as médias dos tratamentos ajustados (Figura 10.4D).

TERMO DE INTERAÇÃO SIGNIFICATIVO Este caso representa inclinações de regressão heterogêneas, fazendo-se necessário ajustar uma linha de regressão separada, com sua própria inclinação e intercepto, para cada grupo de tratamento (Figura 10.4E,F). Quando a interação de tratamento × covariável é significativa, pode não ser possível discutir um efeito geral do tratamento, porque a diferença entre tratamentos pode depender do valor da covariável. Se a interação é forte e as linhas das regressões se cruzarem, a ordem das médias dos tratamentos será invertida nos valores altos e baixos da covariável (Figura 10.4E). Se o efeito do tratamento



é forte, a ordem das médias dos tratamentos pode permanecer a mesma para valores diferentes da covariável, mesmo que o termo de interação seja significativo (Figura 10.4F).

COMPARANDO MÉDIAS

Na seção anterior, enfatizamos que comparar as médias de diferentes grupos de tratamentos é essencial para interpretar corretamente os resultados da análise de variância. Mas como decidimos quais médias são verdadeiramente diferentes

Figura 10.4 Possíveis resultados para experimentos com delineamentos de ANCOVA. Em cada painel, símbolos diferentes indicam réplicas em cada um dos três grupos tratamentos. Cada painel indica um diferente resultado experimental possível com o resultado da ANCOVA associado. (A) Efeitos do tratamento e da covariável não significativos. Os dados são melhor ajustados por uma única média geral com erro amostral. (B) Efeito da covariável significativo, efeito do tratamento não significativo. Os dados são melhor ajustados por uma única linha de regressão sem diferenças entre tratamentos. (C) Efeito do tratamento significativo, covariável não significativa. O termo da covariável não é significativo (inclinação da regressão = 0) e os dados são melhor ajustados por um modelo com médias de grupos diferentes, como em uma ANOVA de um fator simples. (D) O tratamento e a covariável são significativos. Os dados são melhor ajustados por um modelo de regressão com uma inclinação comum, mas interceptos diferentes para cada grupo de tratamento. Este modelo é utilizado para calcular médias ajustadas, estimadas para a média geral da covariável. (E) A interação tratamento × covariável em que a ordem das médias dos tratamentos difere para diferentes valores da covariável. As inclinações da regressão são heterogêneas e cada grupo de tratamento possui uma inclinação e intercepto diferentes. (F) Uma interação tratamento × covariável na qual a ordem dos grupos de tratamentos não difere para os diferentes valores da covariável. Tanto o tratamento quanto a interação tratamento × covariável são significativos.

umas das outras? A ANOVA testa apenas a hipótese nula de que as médias dos tratamentos foram todas amostradas a partir da mesma distribuição. Se rejeitarmos essa hipótese nula, os resultados não especificam quais médias em particular diferem uma da outra.

Para comparar diferentes médias, há duas maneiras. Com comparações *a posteriori* ("depois do fato"), utilizamos um teste em todos os pares possíveis de médias dos tratamentos para determinar quais são diferentes um do outro. Com contrastes *a priori* ("antes do fato"), especificamos antecipadamente as combinações em particular de médias que desejamos testar. Tais combinações com frequência refletem hipóteses específicas que estamos interessados em avaliar.

Embora os contrastes *a priori* geralmente não sejam utilizados por ecólogos e cientistas da área ambiental, somos favoráveis por duas razões: primeiro, como eles são mais específicos, em geral são mais poderosos que testes generalizados de diferenças entre todos os pares de médias. Segundo, o uso de contrastes *a priori* força o investigador a pensar com clareza acerca de quais diferenças de tratamentos, em particular, são de interesse e como elas estão relacionadas com as hipóteses a serem abordadas.

Ilustramos tanto as abordagens *a priori* quanto as *a posteriori* com a análise de um simples delineamento de ANOVA. Os dados são de Ellison et al. (1996), que estudaram a interação entre as raízes do Mangue-vermelho (*Rhizophora mangle*) e esponjas epibentônicas. O mangue está entre as poucas plantas vasculares que podem crescer em água salgada, sendo que, pântanos tropicais protegidos, formam florestas densas, margeando a costa. As raízes-escoras do mangue se estendem para baixo no substrato e são colonizadas por numerosas espécies de esponjas, cracas, algas, pequenos invertebrados e micróbios. A assembleia animal obviamente se beneficia desse substrato duro disponível, mas existe algum efeito sobre as plantas?

Ellison et al. (1996) queriam determinar experimentalmente se havia efeitos positivos de duas espécies comuns de esponjas no crescimento da raiz de

Rhizophora mangle. Eles estabeleceram quatro tratamentos, com 14 a 21 réplicas em cada.[9] Os tratamentos foram os seguintes: (1) sem manipulação; (2) espuma grudada em raízes descobertas de mangue (trata-se de um controle de "esponja falsa" que simula a hidrodinâmica e outros efeitos físicos de uma esponja viva, mas é biologicamente inerte[10]); (3) colônias vivas de *Tedania ignis* (esponja de fogo) transplantadas para raízes descobertas de mangue; (4) colônias vivas de *Haliclona implexiforms* (esponja púrpura) transplantadas para raízes descobertas de mangue. As réplicas foram estabelecidas pré-selecionando de maneira aleatória raízes descobertas de mangue. Os tratamentos foram desordenadamente atribuídos para as réplicas. A variável resposta foi o crescimento da raiz do mangue, medido em mm/dia para cada uma das réplicas. A Tabela 10.12 apresenta a análise de variância para esses dados e a Figura 10.5 ilustra os padrões entre as médias. O valor de P para a razão-F é altamente significativo ($F_{3,52} = 5{,}286$, $P = 0{,}001$) e os tratamentos explicam 19% da variação nos dados (cálculo feito utilizando a Equação 10.23). O próximo passo é focar em quais médias de tratamentos em particular são diferentes umas das outras.

Comparações *a posteriori*

Utilizaremos o teste de "diferença honestamente significante" de Tukey (DHS; *Honestly Significant Difference*) para comparar as médias dos quatro tratamentos no experimento de crescimento do mangue. Este é um de muitos procedimentos *a posteriori* disponíveis para comparar pares de médias após uma ANOVA.[11]

[9] O delineamento completo, na verdade, é um bloco aleatorizado com tamanhos amostrais desiguais, mas o trataremos com uma ANOVA simples de um fator para ilustrar as comparações de médias. Para o nosso exemplo simplificado, analisamos apenas 14 réplicas aleatoriamente escolhidas para cada tratamento. Portanto, nossa análise é completamente balanceada.

[10] O controle de espuma em particular merece atenção especial. Neste livro, discutimos controles experimentais e tratamentos falsos que consideram os efeitos de manipulação e de outros artefatos experimentais que queremos distinguir dos efeitos biológicos significativos. Neste caso, o controle de espuma é utilizado de uma maneira diferenciada. Ele controla os efeitos físicos (p. ex., hidrodinâmica) do corpo de uma esponja que perturba o fluxo da corrente e potencialmente afeta o crescimento das raízes de mangue. Portanto, o controle de espuma não necessariamente controla qualquer artefato de manipulação, porque as réplicas em todos os quatro tratamentos foram medidas e manipuladas da mesma maneira. Em vez disso, esse tratamento nos permite fazer a partição dos efeitos de colônias de esponjas vivas (que secretam e apanham uma variedade de minerais e compostos orgânicos) dos efeitos da fixação de uma estrutura em forma de esponja (que é biologicamente inerte, mas ainda assim altera a hidrodinâmica).

[11] Day & Quinn (1989) discutem extensamente os diferentes testes posteriores à ANOVA e seus pontos fortes e fracos. Nenhuma das alternativas é inteiramente satisfatória. Alguns sofrem muito com erros Tipo I e Tipo II e outros são sensíveis a tamanhos amostrais desiguais ou a diferenças na variância entre grupos. A DHS de Tukey controla as comparações múltiplas, embora seja ligeiramente conservadora com um potencial risco de erro Tipo II (não rejeitar a H_0 quando ela é falsa). *Ver*, também, Quinn & Keough (2002) para uma discussão das diferentes escolhas.

TABELA 10.12 Tabela da análise de variância para um experimento testando os efeitos de esponjas no crescimento de raízes de Mangue-vermelho

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)	Razão-F	Valor de *P*
Tratamentos	3	2,602	0,867	5,286	0,003
Resíduos	52	8,551	0,164		
Total	55	11,153			

Nesta análise, quatro tratamentos foram estabelecidos (sem manipulação, espuma e dois tratamentos com esponjas vivas) contando com 14 réplicas cada (tamanho amostral total = 14 × 4 = 56). A razão-F para o efeito do tratamento é bastante significativa, indicando que as taxas de crescimento médio das raízes do mangue diferem entre os quatro tratamentos. As médias e os desvios-padrão do crescimento da raiz (mm/dia) para cada tratamento são ilustrados na Figura 10.6. (Dados de Ellison et al., 1996.)

O teste DHS de Tukey controla estatisticamente o fato de estarmos realizando muitas comparações simultâneas. Portanto, o valor de *P* deve ser ajustado para baixo em cada teste individual para obter uma taxa de erro sensata do experimento de $\alpha = 0,05$. Na próxima seção, discutiremos em mais detalhes o problema geral de como lidar com valores múltiplos de *P* em um estudo. O primeiro passo é calcular o DHS:

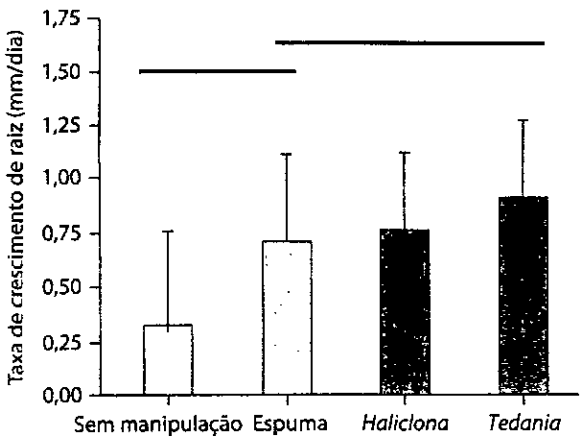


Figura 10.5 Taxa de crescimento médio (mm/dia) de raízes de mangue em quatro tratamentos experimentais. Os tratamentos foram estabelecidos (sem manipulação, espuma e dois tratamentos com esponjas vivas), com 14 réplicas em cada um. A altura da barra é a média da taxa de crescimento para o tratamento e as barras de erro verticais são de um desvio-padrão ao redor da média. As barras em verde-escuro indicam esponjas vivas, a verde-clara é a espuma inerte e a barra não sombreada é a média para as raízes sem manipulação. As linhas horizontais juntam os grupos tratamentos que não diferem significativamente pelo teste de DSH de Tukey (*ver* Tabela 10.13). (Dados de Ellison et al., 1996. As análises estatísticas desses dados estão presentes nas Tabelas 10.12 a 10.15.)

$$DHS = q\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) QM_{resíduo}} \qquad (10.27)$$

onde q é o valor tirado de uma tabela estatística da distribuição de amplitude normalizada (*studentized range distribution*), n_i e n_j são os tamanhos amostrais para as médias dos grupos i e j sob comparação e $QM_{resíduo}$ é o habitual quadrado médio do resíduo da tabela de ANOVA de um fator. Para os dados na Tabela 10.12: q = 3,42 (retirado de um conjunto de tabelas estatísticas), n_i e n_j = 14 réplicas para todas as médias, $QM_{resíduo}$ = 0,164 e o DHS calculado = 0,523.

Portanto, qualquer par de médias entre os quatro tratamentos que diferem por pelo menos 0,523 em crescimento diário médio da raiz é significativamente diferente a um P = 0,05. A Tabela 10.13 mostra a matriz de diferenças par-a-par entre cada uma das médias e os valores de P correspondentes calculados pelo teste de Tukey.

A análise sugere que as raízes sem manipulação diferem bastante do crescimento de raízes com esponjas implantadas. Os dois tratamentos com esponjas vivas não diferem tanto um do outro ou do tratamento com espuma. Entretanto, esse tipo de tratamento foi marginalmente não significativo (P = 0,07) se comparado com as raízes sem manipulação. Esses padrões são ilustrados graficamente na Figura 10.5, na qual as linhas horizontais são colocadas sobre os conjuntos de médias dos tratamentos que não diferem de maneira significativa uns dos outros.

Embora esses testes par-a-par revelem os pares de tratamentos em particular que diferem, o teste DHS de Tukey e outros testes *a posteriori* podem indicar, por vezes, que nenhum dos pares de médias são significativamente diferentes uns dos outros, mesmo que a razão-F total o leve a rejeitar a hipótese nula! Esse resultado inconsistente pode surgir porque os testes par-a-par não são tão poderosos quanto a razão-F global é por si só.

Contrastes *a priori*

Os contrastes *a priori* são muito mais poderosos, tanto estatística quanto logicamente, que os testes de todos os pares *a posteriori*. A ideia é estabelecer **contrastes**, ou comparações especificadas entre conjuntos de médias em particular que testam hipóteses específicas. Se seguirmos certas regras matemáticas, um conjunto de contrastes será ortogonal ou independente dos outros e irá realmente representar a partição matemática da soma dos quadrados entre grupos.

Para criar um contraste, primeiro atribuímos um número inteiro (positivo, negativo ou 0) para cada grupo de tratamento. Este conjunto de coeficientes inteiros é o contraste em teste. Aqui estão as regras para construir contrastes:

TABELA 10.13 Diferenças médias entre todos os possíveis pares de tratamentos em um estudo experimental dos efeitos de esponjas sobre o crescimento de raízes do Mangue-vermelho

	Sem manipulação	Espuma	Haliclona	Tedania
Sem manipulação	0,000			
Espuma	0,383 (0,072)	0,000		
Haliclona	0,436 (0,031)	0,053 (0,986)	0,000	
Tedania	0,584 (0,002)	0,202 (0,557)	0,149 (0,766)	0,000

Nesta análise, quatro tratamentos foram estabelecidos (sem manipulação, espuma e dois tratamentos com esponjas vivas), com 14 réplicas cada (tamanho amostral total = 14 × 4 = 56). A ANOVA é indicada na Tabela 10.12. Essa tabela ilustra as diferenças entre as médias dos tratamentos. Cada valor é a diferença em crescimento médio das raízes (mm/dia) entre um par de tratamentos. A diagonal da matriz compara cada tratamento com ele mesmo, portanto as diferenças são sempre zero. O valor entre parênteses é a probabilidade de cauda associada com base no teste DHS de Tukey. Esse teste controla o nível de α para levar em conta o fato de que seis testes par-a-par foram realizados. Quanto maior a diferença entre os pares de médias, menor é o valor de P. As raízes sem manipulação diferem bastante dos dois tratamentos com esponjas vivas (*Tedania*, *Haliclona*). As diferenças entre todos os outros pares de médias não são significativas. Entretanto, a comparação entre os tratamentos sem manipulação e com espuma é apenas marginalmente acima do nível de $P = 0,05$, enquanto o com espuma não difere de maneira significativa de nenhum dos tratamentos com esponjas vivas. Esses padrões estão ilustrados na forma gráfica na Figura 10.5. (Dados de Ellison et al., 1996.)

1. A soma dos coeficientes para um contraste em particular deve ser igual a 0.
2. Aos grupos de médias que precisam ser calculadas em conjunto são atribuídos o mesmo coeficiente.
3. As médias que não estão incluídas na comparação de um contraste em particular são atribuídas um coeficiente 0.

Vamos tentar isto para o experimento com mangue. Se o tecido de esponjas vivas aumenta o crescimento das raízes, então a média de crescimento nos dois tratamentos com esponjas vivas deveria ser maior que o das raízes no tratamento da espuma inerte.

Nossos valores de contraste são:

Contrate I controle (0) espuma (2) *Tedania* (-1) *Haliclona* (-1)

Haliclona e *Tedania* recebem o mesmo coeficiente (-1) porque queremos comparar a média desses dois tratamentos com o da espuma. O controle recebe um coeficiente 0 porque estamos testando uma hipótese sobre os efeitos de esponjas *vivas*, portanto a comparação relevante é apenas com a espuma. A espuma recebe um coeficiente 2 de forma que ela seja balanceada pela média das duas

esponjas vivas e a soma dos coeficientes seja 0. Equivalentemente, você poderia ter atribuído coeficientes (0) para o controle, (6) para espuma, (–3) para *Tedania* e (–3) para *Haliclona* e, por fim, obter o mesmo contraste.

Uma vez que o contraste é estabelecido, construímos um novo quadrado médio com grau de liberdade associado. Você pode pensar nisto como um quadrado médio ponderado, no qual os pesos são os coeficientes que refletem a hipótese. O quadrado médio para cada contraste é calculado como:

$$QM_{contraste} = \frac{n\left(\sum_{i=1}^{a} c_i \overline{Y}_i\right)^2}{\sum_{i=1}^{a} c_i^2} \qquad \textbf{(10.28)}$$

O termo entre parênteses é apenas a soma de cada coeficiente (c_i) multiplicada pela média do grupo correspondente ($\overline{Y}_i$). Para o nosso primeiro contraste, o quadrado médio da espuma *versus* esponja viva é:

$$QM_{espuma\ vs\ esponjas} = \frac{\left((0)(0{,}3293)+(2)0{,}7121)+(-1)(0{,}7650)+(-1)(0{,}9136)\right)^2 \times 14}{0^2+2^2+(-1)^2+(-1)^2} = 0{,}1510 \quad \textbf{(10.29)}$$

Esta média quadrada possui 1 grau de liberdade e foi testada contra o quadrado médio do erro. A razão-F é 0,151/0,164 = 0,921, com um valor de *P* associado de 0,342 (Tabela 10.14). Este contraste sugere que a taxa de crescimento das raízes de mangue cobertas com espuma é comparável às taxas de crescimento das raízes cobertas com esponjas vivas. Este resultado faz sentido porque as médias desses três grupos de tratamentos são bastante similares.

Não há limite para o número de contrastes em potencial que podemos criar. Entretanto, gostaríamos que nossos contrastes fossem ortogonais uns aos outros, de forma que os resultados fossem logicamente independentes. Os contrastes ortogonais garantem que os valores de *P* não sejam excessivamente inflados ou correlacionados entre si. Com o intuito de criar contrastes ortogonais, duas regras adicionais devem ser seguidas:

4. Se existem *a* grupos de tratamentos, haveria pelo menos ($a - 1$) contrastes ortogonais criados (embora haja muitos conjuntos possíveis de tais contrastes).
5. Todos os pares de produtos cruzados devem somar 0 (*ver* Apêndice). Em outras palavras, um par de contrastes *Q* e *R* é independente se a soma dos produtos de seus coeficientes c_{Qi} e c_{Ri} é igual a zero:

$$\sum_{i=1}^{a} c_{Qi} c_{Ri} = 0 \qquad \textbf{(10.30)}$$

Construir contrastes ortogonais se assemelha à resolução de palavras cruzadas. Uma vez que o primeiro contraste é estabelecido, ele restringe as possibilidades para os remanescentes. Para os dados de raízes de mangue, podemos construir dois contrastes ortogonais adicionais para juntar ao primeiro. O segundo contraste é:

Contraste II controle (3) espuma (-1) *Tedania* (-1) *Haliclona* (-1)

Este contraste soma zero: ele é ortogonal ao primeiro porque satisfaz a regra dos produtos cruzados (Equação 10.30):

$$(0)(3) + (2)(-1) + (-1)(-1) + (-1)(-1) = 0 \qquad \textbf{(10.31)}$$

O segundo contraste compara o crescimento das raízes no controle com a média do crescimento das com espuma e com as duas esponjas vivas. Esse contraste testa se o aumento no crescimento de raiz reflete propriedades das esponjas vivas ou apenas as consequências físicas de uma estrutura fixada. Esse contraste dá uma razão-F bem grande (14,00), que é muito significativa (Tabela 10.14). Novamente esse resultado faz sentido, pois o crescimento médio das raízes sem manipulação (0,329) é substancialmente menor que o do tratamento com espuma (0,712) ou que qualquer um dos dois tratamentos com esponjas vivas (0,765 e 0,914).

Um último contraste ortogonal pode ser criado:

Contraste III controle (0) espuma (0) *Tedania* (1) *Haliclona* (-1)

Os coeficientes deste contraste novamente somam zero e ele é ortogonal aos dois primeiros, pois a soma dos produtos cruzados é zero (Equação 10.30). O terceiro contraste testa as diferenças na taxa de crescimento das duas espécies de esponjas vivas, mas não avalia a hipótese de mutualismo. Ele é de menor interesse que os outros dois, mas é o único contraste que pode ser criado ortogonal aos dois primeiros. Esse contraste produz uma razão-F de apenas 0,947, com um valor de *P* correspondente de 0,335 (Tabela 10.14). Mais uma vez, este resultado é esperado porque as taxas de crescimento médio das raízes nos dois tratamentos com esponjas vivas são similares (0,765 e 0,914).

Note que a soma dos quadrados para cada um dos três contrastes totaliza para soma dos quadrados dos tratamentos:

$$SQ_{tratamento} = SQ_{contraste\ I} + SQ_{contraste\ II} + SQ_{contraste\ III} \qquad \textbf{(10.32)}$$
$$2{,}602 = 0{,}151 + 2{,}296 + 0{,}155$$

Em outras palavras, acabamos de fazer a partição da soma total dos quadrados entre grupos em três contrastes ortogonais e independentes. Isto é similar ao que acontece em uma ANOVA de dois fatores, na qual fazemos a decomposição da soma global dos quadrados dos tratamentos em dois efeitos principais e um termo de interação.

Este conjunto de três contrastes não é a única possibilidade. Também poderíamos ter especificado esse conjunto de contrastes ortogonais:

Contraste I	controle (1)	espuma (1)	*Tedania* (–1)	*Haliclona* (–1)
Contraste II	controle (1)	espuma (–1)	*Tedania* (0)	*Haliclona* (0)
Contraste III	controle (0)	espuma (0)	*Tedania* (1)	*Haliclona* (–1)

O Contraste I compara o crescimento médio de raízes cobertas com as esponjas vivas com o dos dois tratamentos sem esponjas vivas. O Contraste II compara especificamente o crescimento das raízes do controle *versus* as raízes cobertas com espuma. O contrate III compara taxas de crescimento das raízes cobertas com *Haliclona versus* raízes cobertas com *Tedania*. Estes resultados (Tabela 10.15) sugerem que esponjas vivas incrementam o crescimento das raízes, em comparação às raízes sem manipulação ou àquelas cobertas com espuma. Entretanto, o Contraste II revela que o crescimento das raízes com espuma teve um incremento em comparação com controles sem manipulação, sugerindo que o crescimento de raízes de mangue responde à hidrodinâmica ou a alguma outra propriedade física das esponjas aderidas em vez da presença de tecidos vivos de esponjas.

Qualquer um dos conjuntos de contrastes é aceitável, embora pensemos que o primeiro conjunto endereça mais diretamente os efeitos da hidrodinâmica *versus* as esponjas vivas. Por fim, notamos que pode haver pelo menos dois contrastes não ortogonais interessantes. Eles podem ser:

controle (0)	espuma (1)	*Tedania* (–1)	*Haliclona* (0)
controle (0)	espuma (1)	*Tedania* (0)	*Haliclona* (–1)

Esses contrastes comparam o crescimento de raízes de cada uma das esponjas vivas com o de raízes com o controle de espuma. Ambos não são significativos (*Tedania* $F_{1,52} = 0{,}12$, $P = 0{,}73$; *Haliclona* $F_{1,52} = 0{,}95$, $P = 0{,}33$), novamente sugerindo que o efeito das esponjas vivas sobre o crescimento de raízes de mangue é semelhante ao da espuma biologicamente inerte. Entretanto, esses dois contrastes não são ortogonais um ao outro (ou aos nossos outros contrastes), porque seus produtos cruzados não somam zero. Portanto, esses valores de P calculados não são completamente independentes dos valores de P que calculamos com os outros dois conjuntos de contrastes.

Neste exemplo, as comparações *a posteriori* utilizando o teste DHS de Tukey deram resultados um pouco ambíguos porque não foi possível separar, com clareza, os valores médios com base em comparações par-a-par (Tabela 10.13). Os contrastes *a priori* (em oposição) deram resultados muito mais claros, com uma

TABELA 10.14 Contrastes *a priori* para análise de variância dos efeitos de esponjas no crescimento de raízes de mangue

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)	Razão-F	Valor de *P*
Tratamentos	3	2,602	0,867	5,287	0,026
Espuma *vs.* esponjas	1	0,151	0,151	0,921	0,342
Sem manipulação *vs.* esponjas, espuma	1	2,296	2,296	14,000	<0,001
Tedania vs. Haliclona	1	0,155	0,155	0,945	0,335
Resíduo	52	8,539	0,164		
Total	56	11,141			

Os tratamentos foram estabelecidos conforme as Tabelas 10.12 e 10.13. Acabamos de fazer a partição de um simples teste da ANOVA de um fator dos efeitos do tratamento em três contrastes independentes, cada um com 1 grau de liberdade. Cada contraste é testado contra o quadrado médio dos resíduos. A análise revela que o crescimento de raízes sem manipulação é muito diferente do crescimento médio daquelas cobertas com esponjas vivas ou com espuma. Não houve diferença significativa entre o crescimento de raízes com espuma e com esponjas vivas. O mesmo entre o crescimento de raízes cobertas com as duas espécies diferentes de esponjas (*Tedania* e *Haliclona*). Note que foi feita a partição dos três graus de liberdade e da soma dos quadrados do efeito total dos tratamentos em três componentes aditivos, cada um deles testa uma hipótese *a priori* mais específica acerca das diferenças entre médias. (Dados de Ellison et al., 1996.)

evidente rejeição da hipótese nula e a confirmação de que os efeitos das esponjas podem ser atribuídos aos efeitos físicos da colônia de esponjas, e não necessariamente a qualquer efeito biológico das esponjas vivas (Tabelas 10.14, 10.15). Os contrastes *a priori* tiveram mais sucesso por serem mais específicos e mais poderosos que os testes de comparação par-a-par. Note, também, que análises par-a-par necessitaram de seis testes, enquanto os contrastes utilizam apenas três comparações específicas.

Uma nota de advertência: contrastes *a priori* realmente precisam ser estabelecidos *a priori* – ou seja, *antes* de você ter analisado os padrões das médias dos tratamentos! Se forem estabelecidos agrupando tratamentos com médias similares, os valores de *P* resultantes são inteiramente falsos e a chance de erro Tipo I (rejeitar erroneamente a hipótese nula) é grande. Se você não tiver contrastes verdadeiramente *a priori*, e apenas deseja determinar quais médias são diferentes, deve usar, então, um dos métodos *a posteriori*.

Em resumo, contrastes *a priori* são ferramentas poderosas, porém subutilizadas, que permitem a você comparar conjuntos de médias que são em particular relevantes para suas hipóteses. Eles são fáceis de calcular e também podem ser prontamente aplicados a ANOVAs de dois fatores e em delineamentos mais complexos. De fato, você ainda pode usar contrastes *a priori* para fazer a partição dos efeitos principais e dos termos de interação, como em uma ANOVA de dois fatores-padrão. Somos favoráveis a contrastes *a priori* bem pensados em vez dos numerosos testes de comparações *a posteriori* disponíveis.

TABELA 10.15 Um conjunto alternativo de contrastes *a priori* para a análise de variância dos efeitos de esponjas sobre o crescimento de raízes de mangue

Fonte	Graus de liberdade (gl)	Soma dos quadrados (*SQ*)	Quadrado médio (*QM*)	Razão-F	Valor de *P*
Tratamentos	3	2,602	0,867	5,275	0,003
Sem manipulação, espuma *vs.* esponjas	1	1,425	1,425	8,687	0,005
Sem manipulação *vs.* espuma	1	1,027	1,027	6,261	0,016
Tedania vs. Haliclona	1	0,155	0,155	0,947	0,335
Resíduo	52	8,551	0,164		
Total	56	11,153			

Os tratamentos foram estabelecidos conforme as Tabelas 10.12 e 10.13. Novamente, foi feita a partição de um simples teste de ANOVA de um fator dos efeitos do tratamento (Tabela 10.12) em três contrastes independentes, cada um com 1 grau de liberdade. Cada contraste é testado contra o quadrado médio dos resíduos. Estes contrastes são ligeiramente diferentes daqueles da Tabela 10.14. O primeiro contraste indica que a taxa de crescimento de raízes de mangue foi maior nos tratamentos com esponjas vivas que nos tratamentos com espuma ou sem esponjas. O segundo contraste indica que o crescimento das raízes foi também incrementado pelo tratamento com espuma em comparação ao dos controles. Não houve diferença significativa entre o crescimento de raízes cobertas com as duas espécies de esponjas (*Tedania* e *Haliclona*). Note que foi feita a partição dos 3 graus de liberdade e da soma dos quadrados do efeito total dos tratamentos em 3 componentes aditivos, cada um deles testa uma hipótese *a priori* mais específica acerca das diferenças entre médias. (Dados de Ellison et al., 1996.)

CORREÇÕES DE BONFERRONI E O PROBLEMA DOS TESTES MÚLTIPLOS

Em nossa discussão de contrastes *a priori* e de comparações *a posteriori*, fizemos alusão ao problema de comparações múltiplas. Tanto os contrastes *a priori* quanto o teste DHS de Tukey controlam internamente o número de diferentes comparações estatísticas que são feitas. Mas esse tipo de controle não está presente quando comparamos os resultados de testes múltiplos.

A dificuldade é que quanto mais testes estatísticos forem realizados, mais provável será que um valor de *P* menor que 0,05 seja obtido em um ou mais desses testes, mesmo que os padrões reflitam apenas erro aleatório. Por exemplo, se você conduzir 20 testes estatísticos independentes e a hipótese nula for verdadeira em todos, ainda assim, você esperaria rejeitar a hipótese nula 5% das vezes, o que seria aproximadamente (5%) × 20 = 1 teste estatístico. Portanto, muitos autores defendem que você ajuste o nível aceitável de erro Tipo I (α) quando estiver conduzindo testes múltiplos.

O **método de Bonferroni** é para estabelecer a **taxa de erro por experimento** e dividi-la pelo número de testes para obter um nível α ajustado (α') para cada teste individual:

$$\alpha' = \alpha/k \qquad (10.33)$$

onde α é a taxa de erro por experimento, k é o número de testes comparados e α' é o nível do alfa ajustado para cada teste individual. Desta forma, com 20 testes estatísticos e uma taxa de erro por experimento de 0,05, apenas poderíamos rejeitar a hipótese nula de qualquer teste em particular se $P < \alpha'$, que pela Equação 10.33 = 0,05/20 = 0,0025.

O procedimento de Bonferroni é desnecessariamente conservador, porque não considera os casos em que mais de um teste é simultaneamente rejeitado. Uma correção mais adequada é o **método de Dunn-Sidak**, que é o cálculo correto da probabilidade de rejeitar a hipótese nula pelo menos uma vez, com base em regras simples da probabilidade e da combinatória (Capítulo 2). Se α' é o nível de alfa ajustado, então a probabilidade de não cometer nenhum erro Tipo I em k testes é:

$$P(\text{nenhum erro Tipo I}) = (1 - \alpha')^k \qquad (10.34)$$

A probabilidade de fazer pelo menos um erro Tipo I, que é a taxa por experimento, é, portanto,

$$P(\text{pelo menos um erro Tipo I}) = \alpha = 1 - (1 - \alpha')^k \qquad (10.35)$$

Rearranjando, nos termos do nível do alfa ajustado, para cada teste individual temos:

$$\alpha' = 1 - \left(1 - \alpha\right)^{\frac{1}{k}} \qquad (10.36)$$

Para o exemplo dos 20 testes independentes, com uma taxa de erro por experimento de 0,05, temos um valor crítico de $\alpha' = 1 - (1 - 0{,}05)^{1/20} = 0{,}00256$, ligeiramente maior que a simples correção de Bonferroni de 0,00250.

Esses tipos de correções são muito populares entre os editores e revisores de periódicos e com frequência aparecem em artigos publicados. Entretanto, não favorecemos nenhum desses ajustes do nível de alfa e o encorajamos a não os usar, a menos que sua mão seja forçada por um editor.

Por que estamos argumentando contra um padrão aceito? Aqui estão os problemas em ajustar o α. Primeiro, as análises assumem que os testes são to-

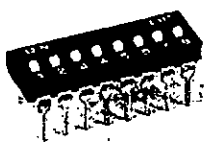
dos independentes, o que raras vezes são. Se os testes não são independentes, o procedimento é muito conservativo (levando a uma alta probabilidade de um erro Tipo II). Segundo, as análises são predicadas na ideia de que todas as hipóteses nulas são verdadeiras, o que mais uma vez representa uma condição extrema.

Mas a objeção mais importante é que o ajuste do α vai contra o senso comum. Uma razão importante para usar o α padrão para rejeitar ou aceitar uma hipótese nula é que o mesmo padrão de evidência é aplicado a todos os testes de hipóteses na literatura científica (Capítulo 4). Porém, se o α for ajustado em cada estudo, o padrão de repente deixa de existir. E qual é precisamente o espaço amostral com que esses ajustes devem ser feitos: o número de testes em um artigo em particular? No fascículo de um periódico como um todo? Ou, talvez, o número de testes realizados durante a vida dos pesquisadores ao longo de suas carreiras? Argumentos legítimos podem ser feitos para cada um desses.

O ajuste efetivo dos α's é uma penalização aos investigadores por conduzirem testes múltiplos, pois o padrão para a rejeição da hipótese nula aumenta quanto mais testes são realizados. No entanto, muitas vezes o padrão com que os testes são rejeitados é a chave para distinguir entre as diferentes hipóteses científicas.[12] O ajuste dos α's nos fazem jogar fora precisamente essa peça chave de informação.

Na verdade, se os testes fossem verdadeiramente independentes, alguém poderia argumentar que os ajustes dos α's na verdade deveriam ser feitos em outra direção! Por exemplo, suponha que conduzimos três experimentos independentes para testar uma hipótese em particular e que o nível de α para rejeitar a hipótese nula em cada caso é de 0,11. Para qualquer teste individual, isso seria algo em torno de 0,05. Porém, com os métodos de Bonferroni ou de Dunn-Sidak, não estamos nem próximos do nível de rejeição necessário (α = 0,0167 e α = 0,0169).

Mas não deveríamos ficar desconfiados? Afinal, se as hipóteses nulas sempre eram verdadeiras, quais são as chances de obter um $P = 0,11$ três vezes em uma linha? A **combinação de probabilidades de Fisher** avalia uma sequência de va-



Interruptor DIP

[12] Rosenzweig & Abramsky (1997) se referem à avaliação de múltiplas peças de evidência como "teste de interruptor DIP". Cada teste representa um interruptor DIP diferente, e o padrão coletivo em que os interruptores são "ligados" fornece uma evidência cumulativa a favor ou contra uma hipótese em particular. Um interruptor DIP, ou mais precisamente um de duplo encapsulamento (*Dual-Inline Package*) – ou seja, com duas posições (ligado e desligado) –, é usado em computadores modernos para configurar o *hardware*. A fotografia ilustra oito interruptores DIP que podem ser arranjados em $2^8 = 256$ combinações de ligado e desligado. Os interruptores DIP substituíram os *jumpers*, que tinham que ser movidos manualmente em placas de circuitos antigas.

lores de probabilidades independentes.[13] O teste estatístico para combinação de probabilidades (CP) é calculado como:

$$CP = -2\sum_{i=1}^{k} \ln(p_i) \qquad (10.37)$$

onde k é o número de testes independentes e $\ln(p_i)$ é o logaritmo natural da probabilidade de cauda para o teste i. A estatística CP segue uma distribuição de chi-quadrado (*ver* Capítulo 11 para uma discussão detalhada de uma distribuição de chi-quadrado) com $2n$ graus de liberdade. Para três testes independentes, cada um com P de 0,11, CP = 13,24, com 6 graus de liberdade. Para os valores tabelados da distribuição de chi-quadrado, a combinação de probabilidades de cauda é 0,039, que podemos declarar como estatisticamente significativa. Isso tem sentido intuitivo, porque em um conjunto de três experimentos de uma distribuição verdadeiramente aleatória seria improvável de gerar três valores de P tão baixos quanto 0,11. Em contraste aos procedimentos de Bonferroni e de Dunn-Sidak, o teste de Fisher será vulnerável ao erro Tipo I se os testes não forem verdadeiramente independentes. Com testes não independentes, é mais provável que você rejeite a H_0 incorretamente, pois cada teste é contado como uma avaliação independente da probabilidade global, o que elas não são.

Com certeza, é importante evitar o excesso de testes estatísticos que não são independentes uns dos outros. Por exemplo, você nunca deve usar uma série de testes-t quando seus dados, na verdade, se ajustam à estrutura de uma ANOVA de um fator com um único valor de P. Por essa mesma razão, preferimos contrastes ortogonais *a priori* às comparações *a posteriori*, que envolvem muitos testes par-a-par que não são independentes uns dos outros. Como veremos no Capítulo 12, a ANOVA multivariada (MANOVA) em geral é preferível a uma série de testes univariados quando todas as variáveis respostas são altamente correlacionadas e medidas nas mesmas réplicas. Similar a isso, a análise de componentes principais e as funções discriminantes muitas vezes são úteis para colapsar um conjunto de variáveis intercorrelacionadas em um número menor de variáveis preditoras ortogonais (Capítulo 12).

Mas uma vez que você tenha reduzido as análises a um número apropriado de testes, nossa preferência é deixar os valores de P brutos e interpretá-los com algum bom senso. A maioria dos revisores e autores seria devidamente cética a um único valor de P significativo dentre um conjunto de 20 testes independentes, enquanto seis ou sete resultados significativos poderiam indicar que alguma coisa interessante está acontecendo. Esperamos que você possa convencer os

[13] Estas diferentes linhas de argumentos também refletem a distinção entre o P de Fisher e o α de Neyman-Pearson (Ver Nota de Rodapé 19, Capítulo 4).

editores e revisores das revistas a deixar os valores de *P* brutos e prestar atenção em como você os interpreta, em vez de constantemente desqualificar todo o seu trabalho em detrimento do uso dos ajustes de Bonferroni.

RESUMO

A análise de variância é uma ferramenta estatística poderosa com base na partição algébrica da soma dos quadrados. Este método assume as amostras aleatórias independentes, as variâncias homogêneas, a distribuição do erro normal, a ausência de erros de classificação e (para certos delineamentos) a aditividade dos efeitos principais. Se esses pressupostos forem cumpridos, as razões-F dos termos de quadrados médios podem ser calculadas testando os efeitos aditivos ou de interação dos diferentes tratamentos. A probabilidade de obter diferentes razões-F pode ser usada para testar estatisticamente a hipótese nula de ausência de efeito dos tratamentos. Do quadrado médio da ANOVA também pode ser feita a partição da variância nos dados em componentes distintos. Diferentes delineamentos experimentais, como o de blocos aleatorizados, de ANOVA aninhada, ANOVA multifatorial, de parcelas subdivididas e ANOVA de medidas repetidas, possuem suas próprias tabelas de ANOVA, nas quais diferentes quadrados médios são utilizados para construir razões-F apropriadas. A análise de covariância (ANCOVA) é uma forma especial de ANOVA, que é um híbrido entre ANOVA e regressão. A construção correta das razões-F é sensível a quando os tratamentos são fatores fixos, em que há apenas alguns níveis de tratamentos distintos, ou fatores aleatórios, cujos níveis do tratamento são um subconjunto aleatório de uma população maior. É importante reconhecer qual delineamento de ANOVA se adequa ao seu delineamento amostral ou experimental. Como muitos delineamentos de ANOVA têm um número equivalente de fatores, é fácil fazer uma análise incorreta, confiando nas configurações pré-estabelecidas dos programas estatísticos. Depois da ANOVA, as médias e as variâncias dos tratamentos podem ser plotadas para revelar os padrões dos efeitos principais e as interações entre os fatores. Testes *a posteriori* podem ser utilizados para fazer comparações par-a-par entre as diferentes médias, mas os contrastes *a priori* são métodos mais poderosos para testar hipóteses e decompor os efeitos dos tratamentos em componentes distintos. As correções de Bonferroni e outros métodos em geral podem ser usados para corrigir os valores de *P* em testes múltiplos, mas bons argumentos podem ser feitos para que valores de *P* sejam interpretados sem nenhum ajuste.

11 Análise de Dados Categóricos

Muitos estudos ecológicos e ambientais geram variáveis respostas que são categóricas, em vez de contínuas. Por exemplo, as plantas podem estar presentes ou ausentes em uma área amostrada, os besouros podem ser vermelhos, alaranjados ou pretos e os tigres forrageando podem virar com mais frequência à direita do que à esquerda. De forma similar, as variáveis preditoras podem ser categóricas em vez de contínuas. Por exemplo, grupos de tratamentos em um estudo experimental ou observacional representam diferentes categorias, como os tipos de esponjas e de raízes de mangues (*ver* Capítulo 10) ou diferentes espécies de besouros. Todos esses são exemplos de **variáveis categóricas**. Os valores que uma variável categórica pode assumir são chamados de **níveis**, podendo ser ordenados (p. ex., pouca luz, luz intermediária, muita luz) ou desordenados (p. ex., besouros, moscas, vespas).

Usamos regressão logística (Capítulo 9) para analisar conjuntos de dados com variáveis preditoras contínuas e variáveis resposta com dois níveis (p. ex., presente, ausente). Usamos ANOVA (Capítulo 10) para analisar conjuntos de dados com variáveis preditoras categóricas e variáveis respostas contínuas. Entretanto, quando *ambas* as variáveis, preditora e resposta, são categóricas, os dados resultam de um delineamento tabular (*ver* Capítulo 7) – e nem a ANOVA nem a regressão são ferramentas analíticas apropriadas.

Neste capítulo, damos enfoque a conjuntos de dados que possuem uma única variável resposta, com dois ou mais níveis (como a presença ou a ausência de um organismo, ou a cor da pelagem de um animal). Os dados de estudos desse tipo representam contagens – ou frequências – de observações em cada categoria. Quando há apenas uma única variável preditora categórica, os dados são organizados em tabelas de contingência de dois fatores, e as hipóteses são testadas com teste de chi-quadrado, teste-*G* ou teste exato de Fisher (para o caso especial de tabelas de contingência 2×2). Quando há muitas variáveis preditoras, os dados são

organizados em tabelas de contingência multifatoriais, que podem ser analisadas usando modelos log lineares ou árvores de classificação. A análise bayesiana de tabelas de contingência fornece estimativas probabilísticas da frequência esperada para cada célula das tabelas de contingência de dois fatores, ou para os parâmetros de modelos log lineares e árvores de classificação. A análise de muitas variáveis categóricas é discutida no Capítulo 12.

Concluímos este capítulo com uma discussão de testes de bondade dos ajustes. Podemos testar quando a amostra de dados se conforma com, se tem consistência com, ou se **ajusta** a uma distribuição conhecida, como a binomial, Poisson ou normal. Os testes de bondade do ajuste – testes chi-quadrado,-*G* e de Kolmogorov-Smirnov – podem ser usados para verificar os resíduos de dados brutos ou de dados transformados que seguem uma distribuição normal, um requerimento de muitos testes paramétricos.

TABELAS DE CONTINGÊNCIA DE DOIS FATORES

Organizando os dados

Os dados categóricos são organizados tipicamente em tabelas de contingência de dois fatores, nas quais as linhas representam os níveis da variável preditora e as colunas, os da variável resposta. Os registros nessas tabelas são as contagens – ou **frequências** – de observações em cada categoria. Eis os dados fundamentais para a análise de tabelas de contingência.[1] A análise de tabelas de contingência é feita corretamente apenas sobre as contagens brutas, não sobre porcentagens, proporções ou frequências relativas dos dados.

Ilustramos testes de tabelas de contingência de dois fatores usando dados de um estudo que examinam os fatores associados ao estado de 73 populações de espécies de plantas raras da Nova Inglaterra (Farnsworth, 2004). Para cada população registrou-se se estava em declínio ou não, se era legalmente protegida, a presença ou ausência de espécies invasoras e cinco níveis qualitativos de luminosidade (de 0 para locais escuros até 4 para os mais claros). As três primeiras linhas desse conjunto de dados são ilustradas na Tabela 11.1. Um resumo da associação entre duas das variáveis categóricas desse conjunto de dados é apresentado na Tabela 11.2 em forma de **tabela de contingência.**

Essa tabela de contingência ilustra a relação entre o estado de proteção das populações de espécies raras e quando ou não as populações estavam em declí-

[1] A terminologia associada aos dados de tabelas de contingência é um pouco inconsistente. Tecnicamente falando, **frequências** são as contagens brutas de observações em cada célula da tabela de contingência. As **frequências relativas** são **proporções** obtidas pela divisão do número de observações de uma célula pelo total da linha, total da coluna ou pelo total geral da tabela. Por fim, as **porcentagens** são proporções que foram multiplicadas por 100, de forma que variam de 0 a 100. Porém, muitos cientistas usam o termo "frequências" para indicar aquelas relativas e os termos "porcentagem" e "frequências" indiferentemente. A distinção entre esses termos é crítica, entretanto, pois os testes estatísticos com dados em tabelas de contingência precisam ser realizados com as contagens brutas (i. e., frequências), não com frequências relativas ou proporções.

TABELA 11.1 Três linhas de um conjunto de dados com medidas de fatores associados com o estado populacional de espécies de plantas raras na Nova Inglaterra

Espécies	Espécie invasora presente?	População em declínio?	Proteção legal?	Nível de luz
Aristolochia	Não	Não	Não	2
Hydrastis	Não	Sim	Não	0
Liatris	Sim	Sim	Não	4
...	...	...	...	...

Cada observação nesse estudo, é de uma população em particular de uma espécie. Para cada população, a tabela de dados registra a presença ou a ausência de espécies invasoras, se a população está em declínio ou não, se está protegida legalmente ou não, e do nível de luminosidade do sub-bosque, que varia de 0 (menor luminosidade) a 4 (maior luminosidade). O tamanho amostral total é de 73 espécies. (Dados de Farnsworth, 2004)

nio. Nesta análise usaremos o estado de proteção como variável preditora e o estado da população como variável resposta. Estabelecemos assim o modelo porque acreditamos que o de proteção afeta potencialmente o estado da população, e não o contrário. Note, contudo, que, ao montar uma tabela de contingência, não importa se você coloca os preditores em linhas e respostas em colunas, ou vice-versa. Os testes estatísticos aplicados a tabelas de contingência dão o mesmo

TABELA 11.2 Tabela de contingência de dois fatores resumindo a relação entre a proteção e o estado da população

	Estado de proteção		
Estado da população	Sem proteção	Protegida	*Total da linha*
Em declínio	$Y_{1,1} = 18$	$Y_{1,2} = 8$	$\sum_{j=1}^{m} Y_{1,j} = 26$
Estável ou aumentando	$Y_{2,1} = 15$	$Y_{2,2} = 32$	$\sum_{j=1}^{m} Y_{2,j} = 47$
Total da coluna	$\sum_{i=1}^{n} Y_{i,1} = 33$	$\sum_{i=1}^{n} Y_{i,2} = 40$	$\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{i,j} = 73$

Cada célula na tabela especifica uma combinação particular de proteção e estado da população. Os números indicam quantas populações foram observadas dentro de uma classificação em particular. Por exemplo, existem 18 populações desprotegidas que estão em declínio e 32 protegidas que estão estáveis ou aumentando. Algumas tabelas de contingência podem ter valores de 0 quando não houver observações para uma combinação particular de fatores. Em qualquer tabela de contingência de dois fatores, a célula com o valor do total geral (73) é igual à do total das linhas (26 + 47), que também é igual à soma do total das colunas (33 + 40). As somas das colunas e das linhas às vezes são chamadas de total marginal dos dados. (Dados de Farnsworth, 2004)

resultado de qualquer jeito. Interpretar as relações de causa e efeito a partir dos resultados desses testes, contudo, requer a distinção entre preditores e respostas.

As tabelas de contingência de dois fatores possuem linhas e colunas. Por convenção, as linhas são indexadas usando a letra i e as colunas, j. Os índices das linhas variam de 1 a n, onde n é o número de categorias nas linhas (duas na Tabela 11.2). Os índices das colunas variam de 1 a m, onde m é o número de categorias das linhas (duas na Tabela 11.2). O valor $Y_{i,j}$ em cada célula representa a frequência do número de observações que estão tanto no nível representado na linha quanto no nível representado na coluna. Por exemplo, o valor 18 na célula superior esquerda ($Y_{1,1}$) da Tabela 11.2 significa que 18 populações foram amostradas e estavam em declínio e desprotegidas.[2]

As outras quantidades que são derivadas de tabelas de contingência correspondem ao **total das linhas**, ao **total das colunas** e ao **total geral**. O total das linhas é a soma das observações em cada linha (na Tabela 11.2, há um total de 26 populações em declínio e 47 estáveis ou aumentando). O total das colunas é a soma das observações em cada coluna (na Tabela 11.2, há um total de 33 populações que estão desprotegidas e 40 que têm certo grau de proteção legal). O total geral (que é igual ao tamanho amostral total) pode ser calculado de três formas: como a soma do total das linhas (26 + 47 = 73), a soma do total das colunas (33 + 40 = 73) ou a soma das frequências em cada célula (18 + 8 + 15 + 32 = 73). Se todas essas somas não forem iguais, confira sua aritmética!

Uma forma conveniente de visualizar os dados em tabelas de contingência é por meio de um **gráfico mosaico** (Friendly, 1994). A Figura 11.1 ilustra um gráfico mosaico dos dados mostrados na Tabela 11.2. Nesse gráfico, os dois "eixos" são as duas variáveis – estado de proteção no eixo-x (a variável preditora) e estado da população no eixo-y (a variável resposta). O tamanho dos "retângulos" é dimensionado na proporção de observações em cada célula da tabela de contingência. Portanto, a área total de cada retângulo é proporcional à frequência relativa das combinações.

As variáveis são independentes

ESPECIFICANDO A HIPÓTESE NULA As tabelas de contingência são usadas para testar a hipótese nula de que as variáveis preditora e resposta não são associadas uma com a outra. No exemplo mostrado na Tabela 11.2, a questão é se há uma associação ou relação entre o estado de proteção da população e se ela está ou

[2] Infelizmente, não existem métodos padrão para organizar tabelas de contingência em planilhas, nem formas padronizadas com as quais os programas estatísticos manipulem dados categóricos. Alguns programas importam tabelas de contingência diretamente, outros geram essas tabelas a partir de conjuntos de dados em que cada linha representa uma única observação, e ainda há os que trabalham dados em ambos os formatos. Se você tiver um pacote estatístico favorito e souber que trabalhará com dados categóricos, confira os requerimentos de formato antes de começar a codificar seus dados.

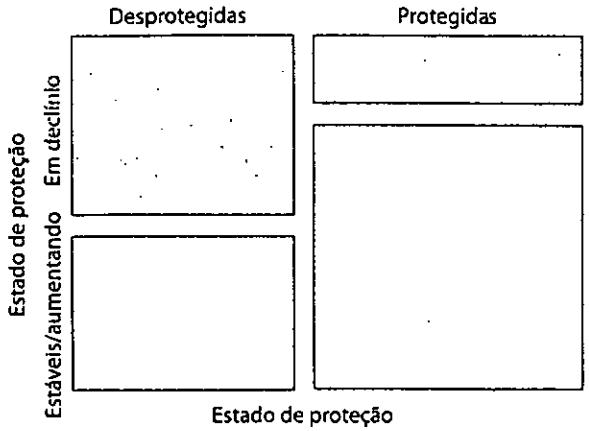


Figura 11.1 Ilustração de um gráfico mosaico. Esse gráfico retrata a relação entre o estado das populações de plantas raras e quando as terras em que elas ocorrem são protegidas ou não (Farnsworth, 2004). As frequências nas células (da Tabela 11.2) são representadas como "retângulos", cujos tamanhos são proporcionais à frequência relativa no conjunto de dados. A largura das colunas é proporcional ao total das colunas. Por exemplo, como 40 das 73 populações são protegidas, a coluna da proteção ocupa 55% (ou 40/73) da largura do gráfico. A altura de cada retângulo é proporcional à frequência da célula. Deste modo, 32 das 40 populações protegidas estão estáveis ou aumentando, portanto seu retângulo ocupa 32/40, ou 80%, da coluna da direita.

não em declínio. Especificamente, precisamos saber se populações desprotegidas são mais prováveis de estarem em declínio do que as protegidas. Se esse padrão puder ser estabelecido, ele pode sugerir que a proteção tem um efeito mensurável na prevenção do declínio das populações. Para testar essa hipótese científica, primeiro precisamos especificar uma hipótese nula estatística apropriada. Nesse caso, a mais simples é que as duas variáveis são independentes uma da outra e que o grau de associação observado não é mais forte que o esperado por chance ou amostragem randômica. Nossa discussão no Capítulo 2, sobre a probabilidade de eventos independentes, nos deu uma estrutura de trabalho para testar esta hipótese nula.

CALCULANDO OS VALORES ESPERADOS Nossa hipótese nula diz que os valores em cada uma das quatro células na Tabela 11.2 são independentes uns dos outros. Perguntamos, portanto: quais seriam os **valores esperados** em cada uma das quatro células se a estabilidade da população e seu estado de proteção forem independentes um do outro? Como um exemplo, pegue a célula superior da esquerda, com o número de populações que estão em declínio e não são protegidas. Qual é o número esperado de populações nesta categoria sob a hipótese nula de independência? O valor esperado é o número total de observações (N) vezes a probabilidade (P) de uma população ser desprotegida *e* estiver em declínio:

$$\hat{Y}_{em\ declínio,\ desprotegida} = N \times P(em\ declínio \cap desprotegida) \qquad (11.1)$$

Relembre que, no Capítulo 1, o símbolo de interseção $\cap$ significa a ocorrência simultânea de dois eventos, *A* e *B*. Também, a partir do Capítulo 1 (Equação 1.3), sabemos que a probabilidade de dois eventos independentes ocorrerem ao mesmo tempo é igual ao produto de suas probabilidades individuais:

$$P(em\ declínio \cap desprotegida) = P(em\ declínio) \times P(desprotegida) \qquad (11.2)$$

Como não temos mais nenhuma informação acerca dessas duas populações que não os dados por si só, estimamos as probabilidades de cada um dos termos à direita, na Equação 11.2, a partir dos totais marginais da Tabela 11.2. Por exemplo, estimamos a probabilidade de uma população estar em declínio como o número total de populações em declínio (26 = o total da primeira linha) dividido pelo número total de populações (73). Essa probabilidade é = 0,356. De forma similar, estimamos a probabilidade de desproteção de uma população usando o total da primeira coluna (33 = total de populações desprotegidas) dividido pelo número total de populações (73). Essa probabilidade é = 0,452. Substituindo esses valores na Equação 11.1, temos:

$$\hat{Y}_{em\ declínio,\ desprotegida} = 73 \times 0{,}356 \times 0{,}452 = 11{,}75 \qquad (11.3)$$

Desta forma, se os estados "proteção" e "declínio" forem independentes um do outro, esperaríamos que aproximadamente 12 das 73 populações amostradas estivessem desprotegidas *e* em declínio.

A forma geral para calcular o valor esperado de uma célula em uma tabela de contingência de dois fatores com $i = 1$ a n linhas e $j = 1$ a m colunas é:

$$\hat{Y}_{i,j} = \frac{total\ da\ linhas \times total\ das\ colunas}{tamanho\ amostral} = \frac{\sum_{j=1}^{m} Y_{i,j} \times \sum_{i=1}^{n} Y_{i,j}}{N} \qquad (11.4)$$

Aplicando a Equação 11.4 a cada uma das células em nosso exemplo (Tabela 11.2), obtemos os valores esperados para cada célula (Tabela 11.3). Como usamos os dados para gerar essas probabilidades, o total das linhas, o total das colunas e o total geral dos valores esperados sempre serão idênticos aos totais observados.

Teste de hipóteses: teste de chi-quadrado de Pearson

O teste de chi-quadrado é o método-padrão para testar a hipótese de que as variáveis em uma tabela de contingência de dois fatores são independentes.

TABELA 11.3 Valores esperados para os dados da Tabela 11.2

	Estado de proteção		
Estado da população	**Sem proteção**	**Protegida**	*Total da linha*
Em declínio	11,75	14,25	26
Estável ou aumentando	21,25	25,75	47
Total da coluna	33	40	*Total geral* = 73

A partir da Tabela 11.2, esses valores esperados resumem a relação entre o estado das populações de plantas raras e se elas estão em terras protegidas ou não. Os valores esperados são calculados sob a hipótese nula de que não há uma associação entre o estado de proteção e o da população. Se os dois fatores são independentes, então a frequência esperada em cada célula é apenas o total das linhas vezes o total das colunas, dividido pelo total geral (Equação 11.4). Note que a soma dos valores esperados produz os mesmos totais de linhas e colunas dos dados originais (Tabela 11.2). (Dados de Farnsworth, 2004)

A ESTATÍSTICA-TESTE A estatística-teste, chamada de **estatística chi-quadrado de Pearson** (Pearson, 1900),[3] é calculada como:

$$X^2_{\text{Pearson}} = \sum_{\textit{todas células}} \frac{(\textit{Observado} - \textit{Esperado})^2}{\textit{Esperado}} \qquad (11.5)$$

Em outras palavras, para cada célula, subtraímos o valor esperado do que foi observado, elevamos esta diferença ao quadrado e dividimos pelo valor esperado. Por fim, somamos esses quocientes. Para os dados na Tabela 11.2 e para os valores esperados na Tabela 11.3, temos:

$$X^2_{\text{Pearson}} = \frac{(18-11{,}75)^2}{11{,}75} + \frac{(8-14{,}25)^2}{14{,}25} + \frac{(15-21{,}25)^2}{21{,}25} + \frac{(32-25{,}75)^2}{25{,}75}$$

$$= 3{,}32 + 2{,}74 + 1{,}83 + 1{,}52$$

$$= 9{,}42$$

Karl Pearson

[3] Karl Pearson (1857-1936) foi um dos fundadores da estatística moderna. Após se graduar na Cambridge University, em 1879, ele mudou-se para a University College, Londres, onde, em 1911, e foi nomeado "Galton Professor" de Eugenia. Além do desenvolvimento do teste de chi-quadrado, ele criou o termo "desvio-padrão" para descrever a $\sqrt{\sigma^2}$, inventou o coeficiente de correlação (r), fez significativos avanços na análise de regressão e foi co-fundador da revista *Biometrika*, que publica contribuições estatísticas com aplicações biológicas. Pearson foi particularmente interessado em estatística de amostras grandes (assintótica). Ele discordou de Fisher sobre a importância relativa da estatística de amostras grandes *versus* amostras pequenas. Tal discordância foi significativa o suficiente para Fisher rejeitar uma oferta de emprego de estatístico-chefe do *Galton Laboratory for National Eugenics*, que foi coordenado por Pearson.

Intuitivamente, você pode ver que a estatística chi-quadrado mede a extensão na qual as frequências observadas diferem das esperadas, sob a hipótese nula de independência. Suponha que as frequências observadas em cada célula sejam exatamente iguais às esperadas. O valor do chi-quadrado seria igual a 0. Quanto mais os valores observados se desviam das frequências esperadas, maior o valor do chi-quadrado. Ainda, note que a estatística do chi-quadrado é uma medida elevada ao quadrado, portanto somente importa o valor do desvio, e não se ele é negativo ou positivo. Neste sentido, a estatística chi-quadrado é bastante similar à soma dos quadrados dos resíduos que descrevemos para a regressão linear (Capítulo 9).

A distribuição esperada da estatística chi-quadrado de Pearson é a distribuição chi-quadrado (χ^2).[4] Como a distribuição-F (Capítulos 5 e 10) e a distribuição-t (Capítulo 3), a forma da distribuição chi-quadrado varia com o número de graus de liberdade.

GRAUS DE LIBERDADE Quais são os graus de liberdade para a Tabela 11.2? Para um dado tamanho amostral (73 neste caso) e para os dois totais das linhas (26 e 47 neste caso) e para os dois totais das colunas (33 e 40), assim que o valor de uma célula seja especificado, todos os outros são fixos. Desta forma, há apenas um grau de liberdade nesta tabela 2×2.

Se a tabela de contingência tiver mais de duas colunas, todos os valores de cada linha podem variar, menos um (uma vez que todos tenham sido especificados, menos um, o valor desse último será determinado pelo total da linha). De forma similar, se a tabela de contingência tiver mais de duas linhas, todas menos os valores de cada coluna podem variar, menos um. Em geral, o número de graus de liberdade de uma tabela de contingência, indicado pela letra grega ν, é igual a:

$$\text{df} = \nu = (\text{número de linhas} - 1) \times (\text{número de colunas} - 1) \qquad (11.6)$$

[4] A distribuição chi-quadrado (χ^2) é uma função densidade de probabilidade que assume valores de = a $+\infty$. Sua função densidade de probabilidade, avaliada apenas para valores positivos de x, é:

$$f(x) = \frac{(\frac{1}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} x^{(\frac{\nu}{2}-1)} e^{-\frac{1}{2}x}$$

onde $\Gamma(x)$ é a função Gamma (*ver* Capítulo 5, Nota de rodapé 11) e ν é o número de graus de liberdade. De fato, uma distribuição χ^2 com ν graus de liberdade é idêntica à distribuição Gamma avaliada com $a = \nu/2$ e $b = 1/2$. A distribuição χ^2 tem apenas um parâmetro, ν. Seu valor esperado = ν, e sua variância esperada = 2ν. A distribuição-F usada em ANOVA (Capítulos 5 e 10) também pode ser expressa como a razão de duas distribuições χ^2.

Esse cálculo dos graus de liberdade difere de nossa experiência com ANOVA (Capítulo 10) e regressão (Capítulo 9). Naquelas análises, os graus de liberdade dependiam, em parte, do tamanho amostral total. Mas, na análise de tabelas de contingência, os graus de liberdade dependem apenas do número de categorias, ou de células, na tabela. Assim, na Tabela 11.2, se tivéssemos dados sobre 730 populações em vez de 73, o teste de chi-quadrado ainda teria apenas 1 grau de liberdade, pois estaríamos analisando uma tabela 2×2. Contudo, o teste estatístico seria muito mais preciso com 730 populações do que com apenas 73. Este ganho em precisão seria perdido se os dados fossem transformados em proporção ou porcentagens, que é uma razão da análise de dados categóricos a ser feita com as frequências brutas.

CALCULANDO O VALOR DE *P* Como em todos os testes de hipóteses, estamos interessados em encontrar a probabilidade dos dados, dada a hipótese nula [$P(\text{dados}|H_0)$], onde a hipótese nula é que as variáveis são independentes. O valor de *P* é a probabilidade de cauda em obter nossos dados ou dados ainda mais extremos, dada a hipótese nula (*ver* Capítulo 4). Para o teste de chi-quadrado de Pearson aplicado à Tabela 11.2, encontramos a probabilidade de obter um valor de X^2_{Pearson} tão grande ou maior que 9.42 (pela Equação 11.5) em relação à distribuição χ^2 com 1 grau de liberdade (pela Equação 11.6). Esse valor pode ser encontrado em uma tabela (p. ex., Rohlf & Sokal, 1995), mas provavelmente ele será fornecido por um programa de estatística. A probabilidade de obter esse valor de $X^2_{\text{Pearson}} = 0{,}0022$. Por ser muito menor que o valor crítico padrão (= probabilidade de cometer um erro Tipo I) de 0,05, rejeitamos a hipótese nula de que as duas variáveis, o estado da população e o estado de proteção são independentes.

Como enfatizamos no Capítulo 4, o valor de *P* é usado para decidir sobre rejeitar ou não a hipótese nula. O próximo passo é examinar um gráfico dos dados e decidir se o padrão estatístico suporta ou contradiz a hipótese científica. Neste caso, a hipótese científica é de que o estado de proteção garante a estabilidade da população, e, de fato, este é o padrão observado na Figura 11.1: das 40 populações protegidas, 80% (= 32) eram estáveis ou estavam aumentando. Em contraste, das 33 populações desprotegidas, apenas 45% (= 15) eram estáveis ou estavam aumentando. A probabilidade do teste de chi-quadrado afirma que essa diferença é improvável de ter ocorrido por chance ($P = 0{,}0022$), e concluímos, a partir dessa simples análise, que o estado de proteção é associado à estabilidade da população.

A estatística do chi-quadrado é o mais familiar e básico teste da independência em uma tabela de contingência. Agora exploraremos duas alternativas para o teste de chi-quadrado: o teste-*G* e o Teste Exato de Fisher.

Uma alternativa ao teste de chi-quadrado de Pearson: o teste-*G*

Como o teste de chi-quadrado, o teste-*G* de independência também compara as frequências observadas com as esperadas. Ele pergunta quão bem a distribuição

dos valores observados se compara à dos valores esperados com base na **distribuição de probabilidades multinomial.**[5]

A ESTATÍSTICA-TESTE A estatística-teste é a razão da probabilidade das frequências observadas com a probabilidade das esperadas (portanto, o teste-*G* muitas vezes é chamado de **teste de razão de verossimilhanças**). Essa razão pode ser calculada de duas formas. A primeira forma, e mais simples, usa os valores esperados que foram calculados com a Equação 11.4 e ilustrados na Tabela 11.3:

$$G = 2 \times \sum_{\textit{todas células}} \left[\text{Observado} \times \ln\left(\frac{\text{Observado}}{\text{Esperado}} \right) \right] \qquad (11.7)$$

Alternativamente, o *G* pode ser calculado diretamente das frequências observadas, sem ter que primeiro calcular os valores esperados:

[5] A distribuição multinomial é uma extensão simples da distribuição binomial (Capítulo 2, Equações 2.2 e 2.3) para variáveis que podem assumir mais de dois valores. Para uma variável aleatória Y que pode assumir j valores, $Y = \{Y_1, \ldots, Y_j\}$, cada um com probabilidade p_j onde pelo primeiro axioma de probabilidade,

$$\sum_{i=1}^{j} p_i = 1$$

e para N tentativas:

$$\sum_{i=1}^{j} Y_i = N$$

a distribuição binomial é definida como:

$$P(Y_1,\ldots,Y_j) = N! \prod_{i=1}^{j} \frac{p_i^{Y_i}}{Y_i!}$$

Para os dados da Tabela 11.2, usando esta fórmula, a probabilidade de observar a frequência das quatro células $Y_{1,1}$, $Y_{1,2}$, $Y_{2,1}$ e $Y_{2,2}$ (que são iguais a 18, 8, 15 e 32, respectivamente) é:

$$\frac{N!}{Y_{1,1}! \times Y_{1,2}! \times Y_{2,1}! \times Y_{2,2}!} \times \left(\frac{Y_{1,1}}{N}\right)^{Y_{1,1}} \times \left(\frac{Y_{1,2}}{N}\right)^{Y_{1,2}} \times \left(\frac{Y_{2,1}}{N}\right)^{Y_{2,1}} \times \left(\frac{Y_{2,2}}{N}\right)^{Y_{2,2}} = 0{,}00202$$

Compare este resultado àquele dado no texto para o teste-*G*.

$$G = 2\times\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\left[Y_{i,j}\ln(Y_{i,j})\right] - \sum_{i=1}^{n}\left[(\sum_{j=1}^{m}Y_j)\ln(\sum_{j=1}^{m}Y_j)\right] - \sum_{j=1}^{m}\left[(\sum_{i=1}^{n}Y_i)\ln(\sum_{i=1}^{n}Y_i)\right] + \left[(\sum_{i=1}^{n}\sum_{j=1}^{m}Y_{i,j})\ln(\sum_{i=1}^{n}\sum_{j=1}^{m}Y_{i,j})\right]\right) \quad (11.8)$$

Em palavras, a Equação 11.8 nos instrui a tomar a soma do produto da frequência de cada célula vezes seu logaritmo natural, subtrair da soma do produto do total das linhas vezes seu logaritmo natural, subtrair da soma do produto do total das colunas vezes seu logaritmo natural e, então, somar o produto do total geral por seu logaritmo natural. Por fim, multiplique tudo por dois. A Equação 11.8 é particularmente útil quando há um grande número de células e é um problema calcular todos os valores esperados. As Equações 11.7 e 11.8 são algebricamente equivalentes, e ambas geram um valor G de 9,575 para os dados na Tabela 11.2.

GRAUS DE LIBERDADE Os graus de liberdade do teste-G são os mesmos para o de chi-quadrado de Pearson, e igual ao produto do (número de linhas – 1) por (número de colunas – 1) (Equação 11.6). Para os dados na Tabela 11.2, há apenas um grau de liberdade.

CALCULANDO O VALOR DE P Como na estatística chi-quadrado de Pearson, a estatística-G é distribuída como uma variável aleatória χ^2 com ν graus de liberdade. O valor de P para $G = 9{,}575$ com 1 grau de liberdade é 0,00197 (dentro do erro de arredondamento do cálculo na Nota de rodapé 5).

O teste de chi-quadrado e o teste-*G* para tabelas *L* × *C*

O teste de chi-quadrado e o teste-G não são restritos a simples tabelas 2×2, como a Tabela 11.2. Eles podem ser usados para qualquer tabela de contingência de dois fatores, independentemente do número de níveis ou de categorias de cada variável. Tais tabelas com frequência são chamadas de tabelas $L \times C$, pois podem ter um número arbitrário de Linhas e Colunas. A Tabela 11.4 apresenta outro subconjunto dos dados de Farnsworth (2004) como uma tabela 2×5 – a relação entre estado da população (duas categorias) e nível de luminosidade (cinco categorias). Para esses dados, a estatística $\chi^2 = 2{,}83$ e tem 4 graus de liberdade [(5 – 1 colunas) × (2 – 1 linhas)], o valor de $P = 0{,}59$. Esse valor é bem maior que 0,05 motivo pelo qual não podemos rejeitar a hipótese nula de que essas variáveis são independentes. De forma similar, a estatística-$G = 3{,}41$, possui 4 graus de liberdade e valor de $P = 0{,}49$, o que também causa a rejeição da hipótese nula de independência. Ambas as análises sugerem que o nível de luminosidade de um local não afeta se uma população está estável ou em declínio. Embora os dados de frequência teoricamente possam ser ajustados a tabelas $L \times C$ de qualquer tamanho, quanto mais categorias houver na tabela, maior deverá ser o tamanho amostral para que o teste funcione de maneira adequada.

TABELA 11.4 Valores observados (em negrito) e esperados (entre parênteses) usados para testar a independência entre o estado da população e o nível de luminosidade

	Estado de proteção					
Estado da população	0	1	2	3	4	*Total da linha*
Em declínio	5 (3,2)	0 (0,7)	3 (3,6)	12 (12,5)	6 (6,0)	26
Estável ou aumentando	4 (5,8)	2 (1,3)	7 (6,4)	23 (22,5)	11 (11,0)	47
Total da coluna	9	2	10	35	17	*Total geral* = 73

Nesta tabela de dois fatores, a associação entre o nível de luminosidade e o estado da população é ilustrado para 73 populações de plantas monitoradas (Tabela 11.1). Os locais foram classificados de acordo com 5 níveis de luminosidade (0 = mais baixa, 4 = mais alta). Embora existam 10 células na tabela, ela ainda é considerada uma tabela de contingência de dois fatores (nível de luminosidade e estado da população). Como na tabela de dois fatores mais simples, os totais marginais e o total geral dos valores observados são iguais aos totais marginais e ao total geral dos valores esperados. Os valores esperados foram calculados sob a hipótese nula de nenhuma associação entre o nível de luminosidade e o estado da população. (Dados de Farnsworth, 2004.)

CORREÇÕES PARA PEQUENOS TAMANHOS AMOSTRAIS E PEQUENOS VALORES ESPERADOS Quando o tamanho amostral total é grande (Legendre & Legendre [1998] consideram "grande" como sendo 10 vezes maior que o número de células na tabela de contingência), tanto a distribuição da estatística chi-quadrado de Pearson quanto a G são igualmente próximas de uma variável aleatória χ^2. Contudo, quando o tamanho amostral total é pequeno – menos de cinco vezes o número de células –, podem haver muitos valores observados iguais a zero ou muitos valores esperados baixos. Os valores esperados baixos constituem um problema em potencial, pois são usados no denominador da estatística chi-quadrado (Equação 11.5). Portanto, se o valor esperado em uma célula for muito pequeno, mesmo um desvio modesto no valor observado pode aumentar muito o valor da estatística chi-quadrado geral. Quando o tamanho amostral é pequeno, a estatística-G deve ser ajustada por meio de um método proposto por Williams (1976):

$$G_{\text{ajustado}} = G/q_{\text{min}} \tag{11.9}$$

onde:

$$q_{\text{min}} = 1 + \frac{\left[N\sum_{i=1}^{n}\frac{1}{\sum_{j=1}^{m}Y_{i,j}} - 1\right] \times \left[N\sum_{j=1}^{m}\frac{1}{\sum_{i=1}^{n}Y_{i,j}} - 1\right]}{6vN} \tag{11.10}$$

Na Equação 11.10, m é o número de colunas, n é o de linhas, N é o tamanho amostral total e ν são os graus de liberdade. Como anteriormente, $Y_{i,j}$ representa a frequência de observações na (linha i, coluna j) da tabela de contingência. O primeiro termo do numerador é 1 menos o tamanho amostral total multiplicado pela soma dos inversos dos totais das colunas. O segundo termo é 1 menos o tamanho amostral total multiplicado pela soma dos inversos dos totais das linhas. Para a Tabela 11.4,

$$q_{\min} = 1 + \frac{[73\times(\frac{1}{9}+\frac{1}{2}+\frac{1}{10}+\frac{1}{35}+\frac{1}{17})-1]\times[73\times(\frac{1}{26}+\frac{1}{47})-1]}{6\times 4\times 73} = 1.109$$

e a estatística-G ajustada = 3,41/1.109 = 3,072, que tem um valor de P de 0,55. Note que a estatística-G ajustada é mais conservativa (gera valores de P maiores) que a G sem ajuste (P = 0,49).

Os livros-padrão de estatística (p. ex., Sokal & Rohlf, 1995) recomendam o uso do G_{ajustado} sempre que qualquer valor esperado for inferior a 5, e alguns programas de estatística produzem mensagens de alerta sobre a validade do chi-quadrado ou teste-G quando os valores esperados são pequenos. Por outro lado, Fienberg (1980) mostrou, em um estudo de simulações, que a correção não é necessária enquanto todos os valores esperados forem maiores que 1. Na prática, contudo, o uso do G_{ajustado} *vs.* o G apenas faz diferença quando o tamanho amostral for pequeno e os valores de P estiverem próximos de 0,05. Apesar das mensagens de aviso, a maioria dos pacotes estatísticos não incorpora o G_{ajustado} em seus menus de opções, embora muitos incluam outra correção, em que 0,5 é somado ou subtraído de cada valor observado na tabela. Essa correção, conhecida como correção de continuidade de Yates (Sokal & Rohlf, 1995) também resulta em valores menores de G ou X^2_{Pearson}, levando, como o G_{ajustado}, a um teste mais conservativo (i. e., um teste menos provável de rejeitar a hipótese nula).

VALORES ASSINTÓTICOS *VERSUS* TESTES EXATOS O teste de chi-quadrado de Pearson e o teste-G são assintóticos – ou seja, a distribuição do G e a do X^2_{Pearson} se aproximam de uma variável aleatória χ^2 conforme o tamanho amostral se torna infinitamente grande. Portanto, os valores de P calculados têm como base uma aproximação, embora razoável. Por meio de um pressuposto-chave, contudo, é possível calcular um valor de P exato para uma tabela de contingência $L \times C$. Esse pressuposto é de que os totais das linhas e das colunas são fixados *a priori* pelo pesquisador.

Por exemplo, Na Tabela 11.2, um teste exato seria apropriado se a intenção do estudo fosse a de amostrar exatamente 26 populações em declínio, 47 estáveis ou aumentando, 33 desprotegidas e 40 protegidas. Embora esse pressuposto seja improvável nesse exemplo, não é difícil imaginar experimentos nos quais o número de indivíduos em cada grupo de tratamento seja fixo e a intensidade do tratamento aplicado aos indivíduos também. Mais tipicamente, em estudos

amostrais, ou o total das linhas ou o das colunas é fixado pelo investigador, mas não ambos. Assim, podíamos ter escolhido amostrar 33 populações desprotegidas e 40 protegidas, mas não ter colocado restrições sobre o estado da população. Ou, poderíamos ter selecionado 26 populações em declínio e 47 estáveis, sem considerar seu estado de proteção.

Quando os totais das linhas e das colunas são fixados, o cálculo de um valor de P exato é, do ponto de vista conceitual, simples, mas, do computacional intensivo. A probabilidade exata (ou valor de P) é a de obter a frequência observada nas células e todas as outras frequências daquelas que são mais extremas que os valores esperados, dado o valor fixo para o total das linhas e das colunas. Para uma tabela 2 × 2, este procedimento é chamado de Teste Exato de Fisher. Como antes, a hipótese nula é de que as variáveis das linhas e das colunas são independentes.

Primeiro, calculamos o número total de permutações (*ver* Capítulo 2) que podem resultar de uma tabela 2 × 2 com os totais de linha e colunas fixos. Este valor é:

$$\binom{N}{Y_{1,1}+Y_{1,2}}\binom{N}{Y_{1,1}+Y_{2,1}} = \frac{N!}{(Y_{1,1}+Y_{1,2})!(Y_{2,1}+Y_{2,2})!} \times \frac{N!}{(Y_{1,1}+Y_{2,1})!(Y_{1,2}+Y_{2,2})!} \tag{11.11}$$

Segundo, a partir da distribuição multinomial (*ver* Nota de rodapé 5, neste capítulo), calculamos o número de formas nas quais podemos obter a exata combinação de valores das células que observamos:

$$\frac{N!}{Y_{1,1}!\times Y_{1,2}!\times Y_{2,1}!\times Y_{2,2}!} \tag{11.12}$$

A divisão da Equação 11.12 pela Equação 11.11 resulta na probabilidade da combinação precisa dos valores das células que observamos (equivalente ao cálculo do número de sucessos dividido pelo número de tentativas):

$$P_{observado} = \frac{(Y_{1,1}+Y_{1,2})!\times(Y_{2,1}+Y_{2,2})!\times(Y_{1,1}+Y_{2,1})!\times(Y_{1,2}+Y_{2,2})!}{Y_{1,1}!\times Y_{1,2}!\times Y_{2,1}!\times Y_{2,2}!\times N!} \tag{11.13}$$

A probabilidade de cauda é igual à Equação 11.13 *mais* as probabilidades de todos os casos mais extremos. Para uma tabela 2 × 2, a probabilidade em cauda é calculada apenas com a enumeração de todos os casos mais extremos, reaplicando a Equação 11.13 a cada caso, e somando os resultados. Note que os totais marginais (os valores do numerador) e o total geral (o valor do N no denominador) permanecem os mesmos em todos os casos extremos (apenas os valores $Y_{i,j}$ indivi-

duais no denominador da Equação 11.13 mudam nos outros casos). O valor exato de P é a soma de todas as interações da Equação 11.13; para a Tabela 11.2, o valor exato de $P = 0{,}0031$. Para a tabela de contingência 2×5, ilustrada na Tabela 11.4, o valor exato de $P = 0{,}60$.

Qual teste escolher?

Todas estas três alternativas – o teste assintótico de chi-quadrado de Pearson, o teste-*G* assintótico e seus testes-exatos em contrapartida – dão resultados um pouco diferentes, especialmente para tamanhos amostrais pequenos. Uma forma comum, porém equivocada, de escolher qual usar é executar as três opções e ficar com a que der o melhor resultado – o menor valor de P. Uma forma mais apropriada de escolher é adequar o delineamento amostral do estudo aos pressupostos do teste.

Três delineamentos diferentes podem gerar tabelas de contingência idênticas, e cada um deles tem um teste associado. Primeiro, nosso estudo poderia ter sido definido pelo tamanho amostral total (73, no caso da Tabela 11.2), sem nenhuma atribuição *a priori* do número de populações (ou indivíduos) a cada categoria. Esse é o pressuposto mais comum em estudos de campo: escolha o tamanho amostral total e deixe as partes caírem onde podem. Sokal & Rohlf (1995) se referem a isto como delineamento Modelo I.

Em um delineamento Modelo II, ou o total das linhas ou o total das colunas, nunca ambos, poderá ser fixado por antecipação. No estudo de Farnsworth (2004), o número de populações protegidas foi fixado em 40 por antecipação e o de desprotegidas, em 33 (talvez por causa de decisões regulatórias feitas com antecedência), mas a fração de populações em declínio podia variar livremente.

Por fim, se o investigador fixar *a priori* tanto o total das linhas quanto o total das colunas, temos um delineamento Modelo III. Os delineamentos Modelo II e Modelo III são mais prováveis em estudos experimentais, enquanto o Modelo I é mais comum em estudos observacionais.

O teste-*G* foi desenvolvido explicitamente para delineamentos Modelo I (tamanho amostral fixo), mas ele pode ser usado tanto para delineamentos de Modelos I como para Modelos II (total de linhas ou colunas fixo). O teste de chi-quadrado de Pearson não foi desenvolvido com nenhum delineamento em mente: conforme o tamanho da amostra seja grande, ele dará resultados virtualmente idênticos ao teste-*G*. Quando o tamanho amostral no Modelo I ou no Modelo II for pequeno, o teste-*G* com a correção de William ou de Yates será apropriado. Os testes exatos são, na verdade, apropriados apenas para os delineamentos Modelo III (tanto o total das linhas quanto o das colunas é fixo).

Por fim, existem simulações de Monte Carlo ou de computador análogas para cada um dos três delineamentos. Em uma simulação de Monte Carlo, não usamos uma distribuição estatística para determinar os valores de P, mas, em vez disso, os estimamos diretamente da amostragem (*ver* Capítulo 5). Por exemplo, uma simulação de Monte Carlo para um delineamento Modelo I atribui randomicamente as 73 populações a cada uma das 4 células da Tabela 11.2. Contudo, a probabili-

dade de cair em cada uma daquelas células é proporcional ao valor esperado de cada uma (Equação 11.4). Desta forma, os totais marginais dos dados simulados serão, em média, os mesmos que os valores dos totais marginais observados. Para cada conjunto de dados simulados, a estatística chi-quadrado padrão será calculada. Mil (ou um número maior) valores simulados serão criados e compararemos o valor de chi-quadrado observado diretamente com a distribuição de valores simulados. Portanto, a probabilidade de cauda estimada pela análise de Monte Carlo é 0,010, qualitativamente similar, embora não seja idêntica aos resultados dos testes-*G* e de chi-quadrado.

TABELAS DE CONTINGÊNCIA MULTIFATORIAIS

A maioria dos dados ecológicos inclui mais de duas variáveis (Tabela 11.1). Portanto, com frequência testamos a independência entre múltiplas variáveis preditoras. Como com as tabelas de contingência de dois fatores, os dados são melhor digitados em planilhas (Tabela 11.1), mas melhor organizados como tabelas de contingência (Tabela 11.5).

Organizando os dados

Uma vez que temos mais de dois fatores, os dados são organizados em tabelas de contingência multidimensionais que não podem ser visualizadas tão facilmente como as bidimensionais. Uma forma de mostrar dados categóricos multidimensionais é com múltiplas tabelas de dois fatores, uma para cada nível dos fato-

TABELA 11.5 Classificação completa dos dados de 73 populações de espécies de plantas raras

Espécie invasora		Ausente					Presente				
Nível de luminosidade		0	1	2	3	4	0	1	2	3	4
Estado da população	Protegida										
Em declínio	Não	2	0	2	2	0	1	0	0	7	4
	Sim	1	0	1	1	0	1	0	0	2	2
Estável ou aumentando	Não	0	0	1	1	4	1	1	2	2	3
	Sim	3	1	0	14	2	0	0	4	6	2

As populações são classificadas de acordo com o estado da população (em declínio ou estável), o estado de proteção (sim ou não), a presença de espécies invasoras (sim ou não) e quanto ao nível de luminosidade (5 níveis: 0 = menor, 4 = maior). Cada célula, na tabela, representa o número de populações encontrado em uma combinação de condições em particular. Por exemplo, há 14 populações protegidas que estão estáveis ou aumentando em locais sem espécies invasoras e nível de luminosidade 3. Há 7 populações desprotegidas em declínio, em locais com espécies invasoras presentes e com nível de luminosidade 3. Esses dados podem ser analisados como uma tabela de contingência de quatro fatores. O estado da população é tratado como variável resposta e o nível de luminosidade, o estado de proteção e a presença de espécies invasoras como variáveis preditoras. (Dados de Farnsworth, 2004.)

res com mais dimensões. Por exemplo, para visualizar os dados de Farnsworth (2004) sobre os cinco níveis de luminosidade, podemos juntar 5 tabelas de dois fatores. Cada tabela dessas poderia mostrar a contagem de proteção e o estado da população sob os diferentes níveis de luminosidade. Por fim, se desejarmos incluir o estado de espécies invasoras como um quarto fator, o conjunto inteiro de tabelas de dois fatores teria que ser duplicado para populações com e sem espécies invasoras.

Um leiaute alternativo para uma tabela de contingência com quatro fatores é mostrado na Tabela 11.5. Esse leiaute ilustra a associação entre o estado da população, a presença de espécies invasoras, o estado de proteção e o nível de luminosidade. Cada célula dá o número de populações registrado para uma dada combinação de categorias. Por exemplo, 14 populações protegidas que não estavam em declínio foram registradas em locais com nível de luminosidade 3 e sem espécies invasoras. Sete populações desprotegidas e em declínio foram registradas em locais com nível de luminosidade 3 e com espécies invasoras. Essa tabela é mais difícil de interpretar do que uma de dois fatores, mas um gráfico mosaico (Figura 11.2) novamente fornece uma forma conveniente e rápida para visualizar dados multidimensionais.

Você pode ver que quanto mais fatores são incluídos na tabela, mais específica e detalhada será a caracterização das populações. Portanto, quanto mais fatores são incluídos, mais o tamanho amostral em cada célula é reduzido, pois o número total de observações é dividido entre um número maior de células.

AS VARIÁVEIS SÃO INDEPENDENTES? Como na análise de uma tabela de dois fatores, queremos saber se as variáveis em uma tabela multifatorial são associadas umas com as outras. Contudo, existem muito mais formas para os dados multifatoriais estarem associados uns com os outros do que há para com dois fatores. Antes de passarmos para a rígida especificação de uma hipótese nula adequada, precisamos entender como calcular os valores esperados em uma tabela de contingência multifatorial.

REPRISE: VALORES ESPERADOS PARA UMA TABELA DE CONTINGÊNCIA DE DOIS FATORES Podemos calcular os valores esperados para uma tabela de contingência multifatorial expandindo um método alternativo empregado no cálculo de valores esperados de uma tabela de contingência de dois fatores. Fique atento à nossa descrição desses cálculos, pois irão fornecer não só uma nova forma de testar hipóteses em tabelas dos fatores, mas também uma base firme para analisar hipóteses em tabelas multifatoriais.

Para uma tabela de dois fatores com $i = 1$ a n linhas e $j = 1$ a m colunas, calculamos o valor esperado para cada célula usando a Equação 11.4:

$$\hat{Y}_{i,j} = \frac{\textit{total das linhas} \times \textit{total da colunas}}{\textit{tamanho amostral}} = \frac{\sum_{j=1}^{m} Y_{i,j} \times \sum_{i=1}^{n} Y_{i,j}}{N}$$

Se tirarmos o logaritmo natural de ambos os lados da Equação 11.4, obtemos:

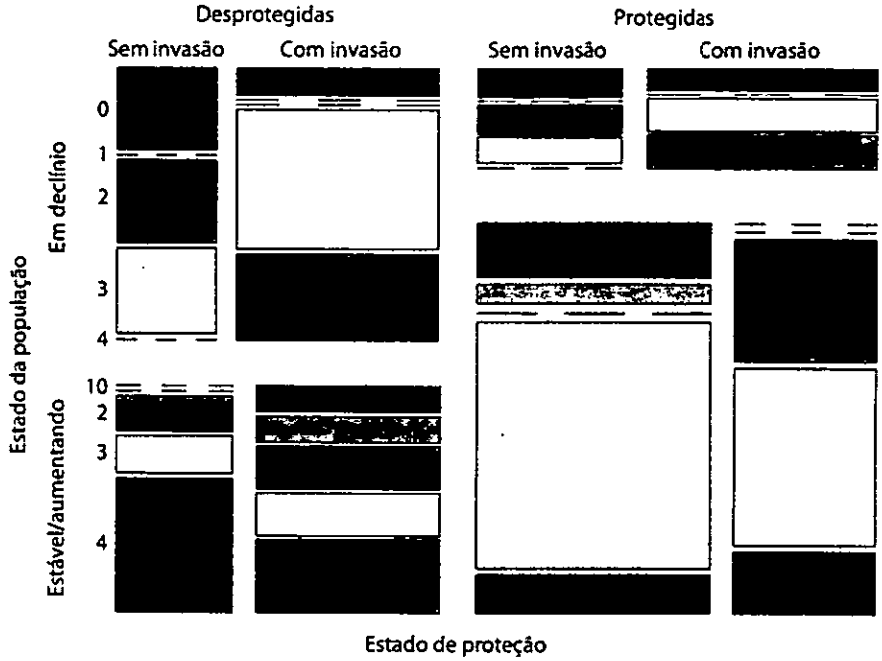


Figura 11.2 Um gráfico mosaico ilustrando os dados da frequência de quatro fatores da Tabela 11.5. Esses dados são para 73 populações de espécies de plantas raras, classificadas de acordo com seu estado de proteção legal (protegida ou não), estado da população (em declínio *versus* estável ou aumentando), presença de espécies invasoras (sim ou não) e intensidade luminosidade (cinco níveis crescentes). Como na Figura 11.1, o tamanho dos retângulos é dimensionado em relação à frequência relativa das células. As categorias principais são o estado da população (em declínio ou estável/aumentando, no eixo-*y*) e a proteção legal (estado de proteção no eixo-*x*), e estes retângulos maiores são do mesmo tamanho em relação àqueles da Figura 11.1. As categorias principais são, então, subdivididas em afetadas ou não afetadas por plantas invasoras (divisões sem invasão e com invasão, no eixo-*x*) e em cinco níveis de luminosidade no eixo-*y*. Cada nível é indicado com um sombreamento diferente (preto = 0, verde mais escuro = 1, verde médio = 2, verde claro = 3 e cinza = 4). As células com frequências = 0 são indicadas por linhas pontilhadas; por exemplo, não há nenhuma população amostrada que esteja desprotegida e em declínio, em locais sem espécies invasoras e com nível de luminosidade 1 ou 4. Dados de Farnsworth (2004).

$$\ln(\hat{Y}_{i,j}) = \ln(\sum_{j=1}^{m} Y_{i,j}) + \ln(\sum_{i=1}^{n} Y_{i,j}) - \ln(N) \qquad (11.14)$$

Assim, o logaritmo natural dos valores esperados em cada célula será a soma dos logaritmos naturais do total da linha, mais o logaritmo natural do total das colunas, menos o logaritmo natural do total geral (que é igual ao tamanho amostral).

Se reescrevermos a Equação 11.14 simbolicamente como:

$$\ln(\hat{Y}_{i,j}) = [\theta] + [A]_{linha} + [B]_{coluna} \qquad (11.15)$$

ela deverá lembrá-lo da equação básica para o modelo de uma análise de variância com dois fatores (Capítulo 10, Equação 10.10),

$$\hat{Y}_{ij} = \mu + A_i + B_j + AB_{ij}$$

embora o termo de interação (AB_{ij}) não conte em nossa equação para tabelas de contingência. Por analogia com a ANOVA, o [θ] na Equação 11.15 é a média geral dos valores preditos (μ na equação da ANOVA) e [*A*] e [*B*] são os efeitos principais, devido às variáveis das linhas e das colunas (A_i e B_j na equação da ANOVA).

Os termos na Equação 11.5 são calculados como a seguir. A média geral [θ] é a mesma dos logaritmos naturais das frequências esperadas em cada célula:

$$[\theta] = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\ln(\hat{Y}_{i,j}) \qquad (11.16)$$

Para cada linha, o efeito principal $[A]_{linha}$ é a média do logaritmo natural da soma das linhas esperadas, menos a média geral:

$$[A]_{linha} = \left[\frac{1}{m}\sum_{j=1}^{m}\ln(\hat{Y}_{i,j})\right] - [\theta] \qquad (11.17)$$

Para cada coluna, o efeito principal $[B]_{coluna}$ é a média do logaritmo natural da soma das colunas esperada menos a média geral:

$$[B]_{coluna} = \left[\frac{1}{n}\sum_{i=1}^{n}\ln(\hat{Y}_{i,j})\right] - [\theta] \qquad (11.18)$$

Note que tanto $[A]_{linha}$ quanto $[B]_{coluna}$ são resíduos: representam os desvios da média geral [θ]. Portanto, se pegarmos a soma de todas as linhas de [*A*] e a soma de todas as linhas de [*B*], os resultados devem ser iguais a 0:

$$\sum_{i=1}^{n}[A]_{linha(i)} = \sum_{j=1}^{m}[B]_{coluna(j)} = 0 \qquad (11.19)$$

A Equação 11.15 não inclui o termo de interação [*AB*] (equivalente ao termo AB_{ij} da equação da ANOVA). Por que não? Em uma tabela de dois fatores,

estamos testando a hipótese nula de que duas variáveis são *independentes* e o termo de interação $[AB] = 0$; nossa hipótese é de que não há interação entre variáveis independentes. Portanto, calculamos os valores esperados para cada célula usando a Equação 11.14, pois ela corresponde à hipótese nula de nenhuma interação.

Sobre as tabelas multifatoriais!

Podemos facilmente estender as Equações 11.14 a 11.19 para tabelas de contingência de qualquer dimensão. Em vez de linhas e colunas, iremos generalizar para d (dimensões) e indexar usando subscritos $i, j, k, \ldots$ Cada dimensão pode assumir um número arbitrário de níveis, que iremos indexar usando os subscritos n, m, p e q... Portanto, para nossos dados com quatro dimensões, na Tabela 11.5, precisamos calcular os valores para as quatro dimensões i, j, k, l. Aplicando a notação da Equação 11.15, usamos o modelo:

$$\ln(\hat{Y}_{i,j,k,l}) = [\theta]+[A]+[B]+[C]+[D]+[AB]+[AC]+[AD]+[BC]+ [BD]+[CD]+[ABC]+[ABD]+[ACD]+[BCD] \quad \textbf{(11.20)}$$

(omitimos os subscritos dos termos entre colchetes [...] para dar clareza). Esta equação é chamada de **modelo log-linear**, pois o logaritmo das frequências esperadas é uma função linear das variáveis preditoras. Você viu outro exemplo de um modelo log-linear quando discutimos regressão logística, no Capítulo 9.

Você também pode notar que o termo de interação completo, $[ABCD]$, não aparece na Equação 11.20. Por quê? Se adicionarmos essa interação na Equação 11.20, os valores esperados seriam exatamente iguais aos observados, e então o ajuste seria perfeito! Mas é claro que não estamos interessados nisso, mas, sim, em quão bem os dados podem ser ajustados por um modelo com menos parâmetros quanto possível. Podemos usar a Equação 11.20, portanto, para testar a hipótese de que o termo da interação completa $[ABCD]$ é igual a zero. Calculamos os valores esperados usando a Equação 11.20 e usamos a Equação 11.15, ou a Equação 11.17, para calcular a estatística chi-quadrado ou a estatística-G. Infelizmente, resolver a Equação 11.20 para obter os valores esperados é difícil (Fienberg, 1970).

Usando o programa S-Plus, calculamos os valores esperados para a Tabela 11.5, calculamos a estatística chi-quadrado (Equação 11.15) e seu valor de P associado, e testamos a hipótese de que as quatro variáveis são independentes. A estatística chi-quadrado = 53,32, para 32 graus de liberdade, tem um valor de P de 0,01. Portanto, rejeitamos a hipótese nula e concluímos que as quatro variáveis não são independentes.

Contudo, esse valor de P geral nos diz apenas que, no mínimo, duas variáveis não são independentes, mas não aponta qual delas. Similarmente, quando obtemos um valor de P pequeno para uma ANOVA de um fator, sabemos que as mé-

dias dos grupos de tratamento em geral diferem, mas não sabemos quais médias em particular são diferentes ou similares. Na ANOVA, usamos contrastes *a priori* e comparações *a posteriori* para determinar quais médias são diferentes (*ver* Capítulo 10). Em análises de tabelas de contingência, dois métodos são disponíveis para determinarmos quais variáveis são independentes e quais não são: modelos log-lineares hierárquicos e árvores de classificação.

MODELOS LOG-LINEARES HIERÁRQUICOS A Equação 11.20 é usada para testar a hipótese de que o termo de interação de quatro fatores $[ABCD] = 0$. Mais especificamente, estamos interessados em determinar quais fatores são associados com o estado da população (em declínio ou estável). A variável *Declínio* é resposta, e as outras três (*Proteção, Invasão* e *Luminosidade*) são as variáveis preditoras. Este modelo, ou hipótese inicial, é equivalente à de que cada resultado da variável *declínio* (sim ou não) é independente das três preditoras, *Proteção, Invasão* e *Luminosidade*. Tal modelo seria parecido com:

$$\ln \hat{Y}_{i,j,k,l} = [\theta]+[A]+[B]+[C]+[D]+[AB]+[AC]+[BC]+[ABC] \quad \textbf{(11.21)}$$

onde $[A]$ é o efeito devido à *Proteção*; $[B]$, devido à *Invasão*; $[C]$, devido à *Luminosidade* e $[D]$, ao *Declínio*. Este modelo inclui todos os termos e a interação das três variáveis preditoras, mas *não* inclui qualquer efeito de interação com *declínio* (fator D). Por que não? Porque nossa hipótese nula é de que o valor de *declínio* é independente das variáveis preditoras – ou seja, não há interação entre *declínio* e qualquer uma das outras variáveis.

Para os dados na Tabela 11.5, a estatística chi-quadrado para o modelo descrito na Equação 11.21 tem um valor de P de 0,004, com 19 graus de liberdade. Portanto, rejeitamos a hipótese nula de que o *Declínio* seja independente das três variáveis preditoras. Agora, precisamos examinar modelos que incorporem algumas interações entre *Declínio* e as outras variáveis. Nossa estratégia é adicionar o número mínimo de interações necessárias para ajustar o modelo. Estamos em busca do tipo mais simples que forneça um bom ajuste aos dados e não cause a rejeição da hipótese nula.

Vamos começar adicionando o termo de interação entre *Declínio* e *Proteção* ($[AD]$):

$$\ln \hat{Y}_{i,j,k,l} = [\theta]+[A]+[B]+[C]+[D]+[AB]+[AC]+[BC]+[ABC]+[AD] \quad \textbf{(11.22)}$$

Este modelo tem uma estatística chi-quadrado = 29,6 com 18 graus de liberdade, produzindo um valor de P marginalmente significativo de 0,04. Esse modelo é "menos significativo" que o da Equação 11.21, que é o que estamos interessados aqui (que "não seja significativamente diferente" do modelo hipotético que "ajuste" os dados).

Como decidimos quais termos de interação incluir no modelo e quais excluir? O problema é exatamente análogo ao que enfrentamos na regressão múltipla e na análise de caminhos (Capítulo 9): escolher um subconjunto de variáveis importantes dentro de um conjunto maior de variáveis preditoras em potencial. A dica é encontrar um balanço entre um modelo que faz um bom trabalho na redução dos desvios entre os valores observados e as predições do modelo, mas que não inclua tantas variáveis, de forma que ele fique superparametrizado. As estatísticas com base no critério de informação de Akaike (AIC) são exemplos que fornecem um índice de "ruindade do ajuste" para levar em conta os desvios residuais e o número de parâmetros no modelo (*ver* a discussão no Capítulo 9).

Para os dados das populações de plantas, o ajuste do modelo, de fato, melhorou com a adição do termo de interação [*Declínio* × *Proteção*]. Essa melhora no ajuste é refletida no decréscimo do AIC de 81,81 para 73,61. Em contraste, a adição de outros termos de interação de dois fatores no modelo, [*Declínio* × *Luminosidade*] ou [*Declínio* × *Invasão*], não melhoram o ajuste do modelo.

Por fim, consideramos a adição de termos de interação de três fatores. A adição de qualquer termo com três fatores (outro que não o [*ABC*] na Equação 11.21) resultou em modelos com ajustes mais fracos, como indicado pelo aumento nos valores de AIC. Tais análises sugerem que o declínio das populações é reduzido pelo estado de proteção, mas elas não são afetadas pelo nível de luminosidade ou pela presença de espécies invasoras no local.

Os modelos que testamos nesta análise são **hierárquicos** – os termos do modelo mais simples (que excluem certas interações) são subconjuntos aninhados dos modelos mais complexos (que os incluem). Considerando todas as combinações possíveis e as interações entre as quatro variáveis, existem 74 modelos alternativos, ou hipóteses, que poderiam ser ajustados! Todos esses modelos incluem os efeitos principais, porém cada um também inclui um conjunto diferente de termos de interação entre as variáveis com efeitos principais. Se essas interações são iguais a zero ou não são as hipóteses de interesse. Esses modelos são hierárquicos, porque, se houver um termo de interação de maior ordem (p. ex., um termo de interação com três fatores), então o modelo também incluirá todos os termos de interação de menor ordem (p. ex., os de dois fatores).

GRAUS DE LIBERDADE PARA MODELOS LOG-LINEARES Você deve ter notado que os diferentes modelos que ajustamos aos dados da Tabela 11.5 tinham graus de liberdade diferenciados. Para uma tabela de dois fatores, os graus de liberdade eram iguais ao produto de 1 menos o número de linhas × 1 menos o número de colunas. Para tabelas multifatoriais, o número de graus de liberdade tem base no somatório dos graus de liberdade para os efeitos principais e para todos os termos de interação. Os graus de liberdade para os termos de interação são calculados exatamente da mesma forma que para a ANOVA (*ver* Capítulo

10). Se o Fator A possui a categorias, e o Fator B possui b categorias, os graus de liberdade para o termo da interação $A \times B = (a-1) \times (b-1)$. Para o modelo completo (todos os efeitos principais e interações, exceto a interação entre os quatro fatores $ABCD$), o número de graus de liberdade ν é calculado como:

$$\begin{aligned}\nu = nmpq - [1 + \\ &(n-1)+(m-1)+(p-1)+(q-1)+ \\ &(n-1)(m-1)+(n-1)(p-1)+(n-1)(q-1)+(m-1)(p-1)+(m-1)(q-1)+(p-1)(q-1)+ \\ &(n-1)(m-1)(p-1)+(n-1)(m-1)(q-1)+(n-1)(p-1)(q-1)+(m-1)(p-1)(q-1)]\end{aligned} \tag{11.23}$$

onde n, m, p e q são os números de possíveis níveis das variáveis A, B, C e D, respectivamente. Em nosso exemplo, os Fatores A, B e D possuem 2 níveis (sim ou não) enquanto o Fator C (luminosidade) possui 5 níveis. Se você olhar atentamente para todos esses termos, deve conseguir reconhecer os componentes dos 4 efeitos principais (a segunda linha da Equação 11.23), os 6 termos de interação entre dois fatores (a terceira linha da Equação 11.23) e os 4 termos de interação entre três fatores (última linha na Equação 11.23). Este modelo completo, portanto, teria 4 graus de liberdade. Note que, quanto mais termos incluímos no modelo, menos graus de liberdade sobram para avaliar o ajuste dos dados ao modelo. Quando testamos um modelo hierárquico, perdemos graus de liberdade, pois os do teste, na verdade, refletem a *diferença* entre os graus de liberdade desse modelo e o próximo modelo na hierarquia.

O primeiro modelo que avaliamos, Equação 11.21, teria:

$$\begin{aligned}\nu = nmpq - [1 + \\ &(n-1)+(m-1)+(p-1)+(q-1)+ \\ &(n-1)(m-1)+(n-1)(p-1)+(m-1)(p-1)+ \\ &(n-1)(m-1)(p-1)]\end{aligned} \tag{11.24}$$

graus de liberdade. Os quatro termos na segunda linha são para os efeitos principais, os próximos três termos correspondem às interações entre os pares ($[AB]$, $[AC]$ e $[BC]$, respectivamente) e o último termo é para a interação entre os três fatores $[ABC]$. A solução da Equação 11.24 para as nossas quatro variáveis = 19 graus de liberdade, como mostramos acima. O modelo na Equação 11.22, que adiciona o termo de interação *declínio* × *proteção* ($[AD]$) tem um grau de liberdade a menos $[(n-1)(q-1) = 1]$, ou 18 graus de liberdade. Você pode não conseguir calcular modelos hierárquicos à mão, mas deve pelo menos conseguir reconstruir os graus de liberdade, entender e interpretar os resultados das estatísticas do modelo ajustado.

ÁRVORES DE CLASSIFICAÇÃO **As árvores de classificação** são úteis, tanto para explorar quanto para modelar dependências entre variáveis tabulares ou categóricas. Embora elas não sejam muito usadas na análise estatística de dados ecológicos, a maioria dos ecólogos e biólogos da área ambiental está um tanto familiarizada com um tipo de árvore de classificação – a chave taxonômica, que é usada para identificar a espécie de um organismo.

Em uma chave taxonômica, uma série de decisões binárias são feitas acerca de características morfológicas do espécime (p. ex., folhas *versus* espinhos). Estas forquilhas das chaves taxonômicas levam a distinções mais e mais específicas (p. ex., folhas simples *versus* folhas compostas) e a chave finalmente termina quando não há mais nenhuma ramificação e chegamos esperançosamente a uma identificação correta da espécie (p. ex., *Acer rubrum* – bordô vermelho). Dados tabulares podem ser analisados com base em estatísticas da mesma forma, com a árvore de classificação indicando forquilhas dicotômicas dos dados que refletem partições nas variáveis categóricas. Faremos, aqui, apenas uma breve introdução a essa técnica, os exemplos ecológicos adicionais são fornecidos por De'ath & Fabricus (2000) e detalhes estatísticos podem ser encontrados em Breiman et al. (1984) e Ripley (1996).

As árvores de classificação são construídas a partir de um conjunto de dados categóricos em grupos cada vez menores, nos quais cada divisão depende de uma única variável. Cada uma resulta em dois grupos, e cada grupo será novamente dividido com base em uma única variável. Em uma chave taxonômica, a divisão prossegue até que haja apenas uma espécie no final de cada "galho". Em árvores de classificação, a divisão continua apenas até que nenhuma melhora no ajuste do modelo seja obtida em relação ao número de divisões aplicado.

Raras vezes, apenas uma das divisões descreverá um grupo. Em vez disso, uma divisão resulta em um grupo caracterizado por uma probabilidade relativamente alta de que ocorra em uma ou mais classes (resultados de uma variável em particular) em relação às outras. A divisão é concluída quando o ponto terminal for o mais homogêneo ou consistente possível, dada a incerteza nos dados. Um nó que prediz que todas as réplicas estejam em uma classe é completamente homogêneo (ou puro) e tem um valor de **impureza** igual a zero. Para árvores de classificação, a impureza é fundamentada na proporção p de réplicas em cada uma das n classes.

A impureza aumenta conforme a habilidade preditiva decresce e pode ser calculada com um índice de informação (entropia) familiar aos ecólogos, como índice de Shannon-Weiner:

$$\text{Impureza} = -\sum_{i=1}^{n} p_i \ln(p_i) \qquad (11.25)$$

ou como o índice de Gini:[6]

$$\text{Impureza} = 1 - \sum_{i=1}^{n} p_i^2 \qquad (11.26)$$

A divisão da árvore ocorre pela minimização da Equação 11.25 ou da Equação 11.26 de cada divisão. Minimizar esses índices é, a grosso modo, equivalente a calcular distribuições de probabilidade condicional de uma variável *A* para cada classe da *B*, como descrito por Legendre & Legendre (1998, 230*ff*.). Para qualquer um desses índices, uma amostra completamente pura ($p = 1,0$, com todas as réplicas classificadas em um grupo) gera um valor do índice de zero.

Mais uma vez usamos os dados da Tabela 11.5 para ilustrar esse procedimento. Como na análise de modelos log-lineares, estamos tentando determinar quais fatores melhor predizem quando uma população está ou não em declínio.[7]

Nosso conjunto de dados contém 73 observações: 26 populações estão em declínio e 43 estáveis ou aumentando. A primeira divisão separa os dados em dois grupos: um com 40 populações e outro com 33 (Figura 11.3). Tal divisão fundamenta-se no estado de proteção. Como discutimos, as populações que estão protegidas são muito mais prováveis de estarem estáveis ou aumentando (32 dentre 40 populações) que em declínio (8 dentre 40 populações, Figura

[6] Ambos índices possuem uma história interessante na ecologia e têm sido usados como medidas de diversidade de espécies. O índice de diversidade é calculado tratando p_i como a fração de indivíduos em uma amostra aleatória que representa a espécie *i*. O índice de Shannon-Weiner veio da teoria de informação (Margalef, 1958) e é uma medida da incerteza ou "informação contida" em uma sequência de dígitos binários. A ideia de que o índice de Shannon-Weiner é uma medida da entropia de uma comunidade ou ecossistema (revisado por Goodman, 1975) não é mais justificada (Hurlbert, 1971). O índice de Shannon-Weiner também possui propriedades estatísticas ruins: ele confunde riqueza de espécies e equitabilidade, além de ser muito dependente do tamanho amostral. Todavia, ele continua sendo usado como um índice de diversidade, particularmente em estudos de poluição da água.

O índice de Gini* é menos popular, mas na verdade possui bases estatísticas mais firmes como uma medida de diversidade. Em uma forma modificada, esse índice é equivalente à Probabilidade de Encontro Interespecífico (PEI), de Hurlbert (1971). A PEI mede a probabilidade de que dois indivíduos, escolhidos randomicamente a partir de uma comunidade, representem duas espécies diferentes. Diferente do índice de Shannon-Weiner, a PEI possui uma interpretação estatística simples e é relativamente insensível ao tamanho amostral (Gotelli & Graves, 1996). Magurran (2003) apresenta uma discussão compreensiva dos potenciais, das fraquezas e das propriedades estatísticas de muitos índices de diversidade ecológica.

[7] Para este exemplo, a construção da árvore foi feita usando o pacote RPART do S-Plus, versão 6.1. Os detalhes da computação para produzir árvores de classificação podem ser encontrados em Ripley (1996) e em Venables & Ripley (2002).

* N. de T. Também chamado de "índice de Simpson".

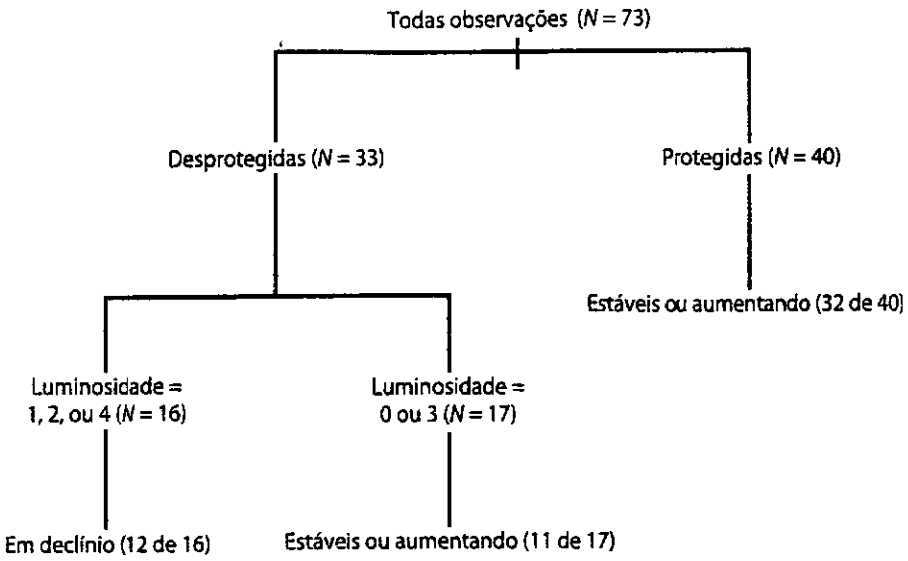


Figura 11.3 Árvore de classificação para os dados de populações de plantas raras da Tabela 11.5. Essa árvore prediz o estado da população (em declínio *versus* estável ou aumentando) em função do de proteção, do nível de luminosidade e se há ou não espécies invasoras presentes. O modelo de melhor ajuste inclui o estado de proteção como forquilha primária, com o nível de luminosidade usado para dividir o grupo de populações desprotegidas em dois grupos. Note que o tamanho amostral final na ponta dos três galhos (16 + 17 + 40) se refere às 73 populações do conjunto de dados (Tabela 11.5).

11.1). Você não deve estar surpreso de que o número total de observações em cada um desses dois grupos seja igual ao total das colunas da Tabela 11.2.

O grupo restante de 33 observações (populações desprotegidas) consiste de 18 populações que estão em declínio e 15 estáveis ou aumentando – quase meio-a-meio. Dividimos esse grupo mais uma vez em dois grupos, agora com base na disponibilidade de luz (Figura 11.3). Dentro do grupo de desprotegidas, as que estão em locais com nível de luminosidade 0 ou 3 são mais prováveis de estarem estáveis ou aumentando (11 de 17 populações), enquanto as que estão em locais com níveis de luminosidade 1, 2 ou 4 provavelmente estão em declínio (12 de 16 populações).

Nenhuma melhora significativa no ajuste é obtida por divisões adicionais desses grupos, portanto a construção da árvore termina aqui. Note que a primeira divisão corresponde à adição do termo de interação *declínio* × *proteção* ([*AD*]) na Equação 11.22. A segunda divisão corresponde à adição do termo *declínio* × *proteção* × *luminosidade* [*ACD*] ao modelo. Na análise log-linear, a inclusão desse termo da interação de três fatores melhorou o ajuste do modelo em relação à Equação 11.22 ($P = 0{,}11$ com esse termo *versus* $P = 0{,}04$ sem ele). Contudo, a penalização por adicionar essa interação (ela reduz os graus de liberdade de 18 para 14) aumentou o AIC de 73,61 para 75,50. Pelas "regras" da seleção de modelos

log-lineares, essa interação de três fatores não foi retida, embora surja como uma forquilha significativa na árvore de classificação.

Não fique incomodado com estas aparentes inconsistências entre os resultados de diferentes tipos de análises estatísticas. Como os modelos log-lineares e as árvores de classificação em diferentes critérios de seleção e diferentes regras de parada, elas podem dar respostas diferentes quando usadas com o mesmo conjunto de dados. Nenhuma análise é "melhor" ou "correta", embora, no contexto de decisões de manejo, a árvore de classificação pode ser mais fácil de interpretar que as análises de modelos log-lineares. As árvores de classificação também podem ser usadas quando os preditores forem variáveis contínuas ou quantitativas. Tais árvores são chamadas de **árvores de regressão**.

Abordagens bayesianas para tabelas de contingência

Os ecólogos com frequência se referem aos testes de chi-quadrado e teste-*G* como não paramétricos. Em contraste, a ANOVA e a regressão são consideradas testes paramétricos, pois elas assumem que os dados e os resíduos têm distribuição normal (*ver* Capítulo 5). Como o teste-*G* e o chi-quadrado geram valores esperados a partir dos próprios dados, alguns pesquisadores erroneamente acreditam que não há nenhuma distribuição de probabilidades subjacente necessária para o teste. De fato, este não é o caso. A modelagem log-linear, subjacente ao teste-*G*, assume que uma distribuição de Poisson ou multinomial descreve os dados. Tal pressuposto também nos permite usar técnicas bayesianas para analisar tabelas de contingência.

Os métodos bayesianos para tabelas de contingência caem em duas categorias amplas. Um conjunto de métodos estima as frequências esperadas como distribuições (com modas ou medianas iguais aos valores esperados). Para dada célula, se o intervalo de credibilidade do valor esperado inclui o valor observado, as linhas e colunas são consideradas independentes (Lindley, 1964; Leonard, 1975; Gelman et al., 1995).

O outro conjunto de métodos estima valores para os parâmetros (com intervalos de credibilidade) para os termos em modelos hierárquicos log-lineares; os parâmetros cujos intervalos de credibilidade incluem o zero são retirados do modelo final. Os métodos bayesianos para escolher entre os modelos log-lineares usam critérios de seleção similares ao AIC (Madigan & Raftery, 1994; Albert, 1996, 1997; Gelman et al., 1995). A aplicação de métodos bayesianos para árvores de classificação é uma área ativa da pesquisa estatística atual (Denison et al., 2002), mas infelizmente ela vai muito além do escopo desse livro de introdução.

TESTES PARA BONDADE DO AJUSTE

Os testes estatísticos que encontramos neste livro, se paramétricos ou bayesianos, assumem que nosso conjunto de amostras ou observações refletem com acurácia

uma variável aleatória com alguma distribuição subjacente. Os exemplos incluem a binomial, Poisson, normal, log-normal ou exponencial (Capítulo 2). As análises consistentes também devem resultar em um conjunto de resíduos que são distribuídos de acordo com uma distribuição conhecida, em geral a normal. Usamos os **testes de bondade do ajuste**, que são similares em estrutura aos de chi-quadrado e *G*, para determinar se nossos dados observados se ajustam a uma distribuição hipotética ou se nossos resíduos se adequam à distribuição normal (ou outra). Os testes de bondade do ajuste também são importantes para testar modelos ecológicos específicos. Se os modelos fazem predições quantitativas acerca dos valores esperados para diferentes categorias de dados, podemos usar testes de bondade do ajuste para ver quão bem nossos dados se adaptam às predições de diferentes modelos (Hilborn & Mangel, 1997).

Um teste de bondade do ajuste funciona quase da mesma forma que os de independência apresentados neste capítulo: um conjunto de observações de uma única variável é comparado aos valores esperados, dada uma distribuição em particular. Os mecanismos são idênticos – use a Equação 11.5, para um teste de bondade do ajuste de um teste de chi-quadrado, ou a Equação 11.7, para um teste de bondade do ajuste de um teste-*G*. Esses dois testes funcionam bem para distribuições discretas e variáveis aleatórias, mas não com distribuições contínuas, como a do tipo normal. Para distribuições contínuas, iremos apresentar um procedimento diferente, o teste de Kolmogorov-Smirnov, para avaliar a bondade do ajuste.

Teste de bondade do ajuste para distribuições discretas

Como um exemplo simples, retornaremos à controvérsia sobre a honestidade da moeda do Euro Belga (*ver* Capítulo 1, Nota de rodapé 7). Dois estatísticos lançaram uma moeda do Euro Belga 250 vezes e observaram 140 caras e 110 coroas. A moeda do Euro Belga é honesta? Em outras palavras, esses dados se ajustam às frequências esperadas de caras e coroas a partir de uma distribuição binomial com $p = 0{,}50$?

TESTE DE CHI-QUADRADO E TESTE–*G* PARA DADOS BINOMIAIS A probabilidade de dar cara em uma moeda honesta é equivalente a 0,5. Portanto, a frequência esperada de caras e coroas são de 125 e 125. Agora, temos valores observados e esperados, e podemos calcular qualquer uma das duas estatísticas: X^2_{Pearson} (Equação 11.5) ou G (Equação 11.7). A estatística $X^2_{\text{Pearson}} = (140 - 125)^2/125 + (110 - 125)^2/125 = 3{,}60$ e a estatística-$G = 2[140 \times \ln(140/125) + 110 \times \ln(110/125)] = 3.61$. Ambos são distribuídos como variáveis aleatórias χ^2 e possuem valores de probabilidade em cauda de 0,0577 e 0,0574, respectivamente. Esses valores de P são muito pequenos, mas, por serem > 0,05, eles não são considerados significativos pelos testes de hipóteses convencionais nas ciências. Portanto, concluímos que os dados observados se ajustam a uma distribuição binomial, e que há evidências insufi-

cientes (apenas fracas) para rejeitar a hipótese nula de que o Euro Belga seja uma moeda honesta.[8]

ALTERNATIVA BAYESIANA Os testes assintóticos clássicos e o teste exato falham, apenas marginalmente, em rejeitar a hipótese nula de que a moeda do Euro Belga é honesta (testando a hipótese $P(\text{dados} \mid H_0)$. Em vez disso, uma análise bayesiana testa $P(H_i \mid \text{dados})$ questionando: os dados realmente fornecem evidência de que o Euro Belga é tendenciosa? Para avaliar essa evidência, precisamos calcular a razão de chances *a posteriori*, como descrito no Capítulo 9 (usando a Equação 9.49). Mackay (2002) e Hamaker (2002) conduziram essas análises usando várias distribuições diferentes de probabilidade *a priori*. Primeiro, se inicialmente não temos nenhuma razão para preferir uma hipótese à outra (H_0: a moeda é honesta; H_1: a moeda é tendenciosa), então a razão de suas priores é = 1 e a razão de chances *a posteriori* é igual à de verossimilhanças

$$\frac{P(\text{data} \mid H_0)}{P(\text{data} \mid H_1)}$$

As verossimilhanças (numerador e denominador) desta razão são calculadas como:

$$P(D \mid H_i) = \int_0^1 P(D \mid p, H_i) P(p \mid H_i) dp \qquad (11.27)$$

onde D são os dados, H_i qualquer uma das hipóteses ($i = 0$ ou 1) e p é a probabilidade de sucesso em uma tentativa binomial. Para a hipótese nula (moeda honesta) a nossa probabilidade *a priori* é de que não há nenhuma tendência. Portanto, $p = 0{,}5$ e a verossimilhança $P(\text{dados} \mid H_0)$ é a probabilidade de obter exatamente 140 caras em 250 tentativas, = 0,0084 (*ver* Nota de rodapé 8). Para a hipótese alternativa (moeda tendenciosa), contudo, existem no mínimo três formas de estipular as priores.

[8] Esta probabilidade também pode ser calculada com exatidão usando a distribuição binomial. A partir da Equação 2.3, a probabilidade de obter *exatamente* 140 caras em 250 lançamentos com uma moeda honesta é:

$$\binom{250}{140} 0{,}5^{140} 0{,}5^{250-140} = 0{,}0084$$

Mas, para um teste de significância bicaudal, precisamos da probabilidade de obter 140 ou mais caras em nossa amostra, mais a probabilidade de obter 110 ou menos coroas em nossa amostra. Por analogia ao Teste Exato de Fisher, discutido neste capítulo, simplesmente adicionamos as probabilidades de obter 140, 141, ..., 250 caras e 0, 1,..., 110 coroas em uma amostra de 250, com probabilidade esperada de 0,50. Esta soma = 0,0581, um valor praticamente idêntico ao obtido com os testes de chi-quadrado e G.

Primeiro, poderíamos usar uma priore uniforme, o que significa dizer que não temos nenhum conhecimento inicial de quanta tendência pode haver; portanto, todas são igualmente prováveis, e $P(p \mid H_1) = 1$. A integral reduz, então, para a soma das probabilidades sobre todas as escolhas de p (de 0 a 1) da expansão binomial para 140 caras entre 250 tentativas:

$$P(D \mid H_1) = \sum_{p=0}^{1} \binom{250}{140} p^{140}(1-p)^{110} = 0{,}00398 \qquad \textbf{(11.28)}$$

A razão de verossimilhança é, portanto, 0,0084/0,00398 =2,1, ou aproximadamente 2:1 chances a favor do Euro Belga ser uma *moeda honesta*!

Alternativamente, poderíamos usar priores mais informativas. A priore conjugada (*ver* Capítulo 5, Nota de rodapé 11) para a distribuição binomial é a distribuição beta,

$$P(p \mid H_1, \alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} p^{\alpha-1}(1-p)^{\alpha-1} \qquad \textbf{(11.29)}$$

onde p é a probabilidade de sucesso, $\Gamma(\alpha)$ é a distribuição gama, e α é um parâmetro variável que expressa nossa convicção *a priori* de que a moeda é tendenciosa. Conforme o α aumenta, nossa convicção na tendência da moeda também aumenta. A verossimilhança de H_1 agora é:

$$P(D \mid H_1) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \binom{250}{140} \int_0^1 p^{140+\alpha-1}(1-p)^{110+\alpha-1} dp \qquad \textbf{(11.30)}$$

Para $\alpha = 1$, temos a priore uniforme (e usamos a Equação 11.28). Resolvendo iterativamente a Equação 11.30, podemos obter razões de verossimilhança para uma ampla gama de valores de α. A razão de chance mais extrema obtida usando a Equação 11.30 (lembre que o numerador é fixo em 0,0084) é 0,52 (quando α = 47,9). Para esta priore informativa, as chances são de aproximadamente 2:1 (= 1/0,52) a favor do Euro Belga ser uma *moeda tendenciosa*.

Por fim, podemos especificar uma priore bem robusta que se adequa exatamente aos dados: $p = 140/250 = 0{,}56$. Agora, a verossimilhança de $H_1 = 0{,}05078$ (a expansão binomial na Equação 11.28, com $p = 0{,}56$) e a razão de verossimilhança é = 0,0084/0,05078 = 0,16, ou 6:1 chances (= 1/0,16) a favor do Euro Belga ter exatamente a tendência observada.

Embora nossas priores mais informativas sugiram que a moeda é tendenciosa, para concluir isso tivemos de especificar priores que sugerem um forte avanço no conhecimento de que o Euro de fato seja tendencioso. Uma análise bayesiana objetiva deixa os dados "falarem por si só" usando priores não informativas, como a uniforme e a Equação 11.28. Aquela análise fornece pouco suporte para a hipótese de que o Euro Belga é tendencioso.

Em soma, as análises assintótica, exata e bayesiana concordam: é mais provável que o Euro Belga seja uma moeda honesta. A análise bayesiana é muito poderosa neste contexto, pois fornece uma medida da probabilidade da hipótese de interesse. Em contraste, a conclusão da análise frequentista recai sobre valores de *P* que são apenas um pouco maiores que o valor crítico de 0,05.

GENERALIZANDO O TESTE DE CHI-QUADRADO E O TESTE-*G* PARA DISTRIBUIÇÕES DISCRETAS MULTINOMIAIS Tanto o chi-quadrado quanto o teste-*G* podem ser usados para avaliar o ajuste de dados que possuem mais de dois níveis de distribuições discretas, como a de Poisson e a multinomial. O procedimento é o mesmo: calcule os valores esperados, aplique a Equação 11.5 ou a 11.7, e calcule a probabilidade de cauda da estatística-teste em relação à distribuição χ^2 com ν graus de liberdade. A única parte complicada é determinar os graus de liberdade.

Distinguimos dois tipos de distribuições que usamos para aplicar testes de bondade do ajuste: distribuições com parâmetros estimados independentemente dos dados e distribuições com parâmetros estimados a partir dos próprios dados. Sokal & Rohlf (1995) chamam-nas de **hipóteses extrínsecas** e **hipóteses intrínsecas**, respectivamente. No primeiro caso, as expectativas são geradas por um modelo ou por uma predição que não depende dos dados. Os exemplos de hipóteses extrínsecas incluem a expectativa 50:50 de uma moeda de Euro honesta, e a razão 3:1 de fenótipos dominantes sobre os recessivos em simples cruzamento mendeliano de dois heterozigotos. Para as hipóteses extrínsecas, os graus de liberdade sempre são um menos o número de níveis que uma variável pode assumir.

As hipóteses intrínsecas são aquelas nas quais os parâmetros precisam ser estimados a partir dos próprios dados. Um exemplo de hipótese intrínseca é ajustar os dados de contagens de eventos raros a uma distribuição de Poisson (*ver* Capítulo 2). Neste caso, o parâmetro da distribuição não é conhecido e precisa ser estimado a partir dos próprios dados. Portanto, usamos um grau de liberdade para o número total de indivíduos amostrados e empregamos outro grau de liberdade para cada parâmetro estimado. Como a distribuição de Poisson possui um parâmetro adicional (λ, a razão de Poisson), os graus de liberdade para este teste seriam dois menos o número de níveis.

Também é possível usar o chi-quadrado ou o teste-*G* para avaliar o ajuste de um conjunto de dados a uma distribuição normal, mas, para fazer isso, os dados devem primeiro ser agrupados em níveis discretos. A distribuição normal possui dois parâmetros a serem estimados a partir dos dados (a média μ e o desvio-padrão σ). Portanto, os graus de liberdade para um teste de bondade do ajuste para o teste-*G* ou de chi-quadrado a uma distribuição normal é igual a três menos o número de níveis. Contudo, a distribuição normal é contínua, e não discreta, e teremos que dividir os dados em um número arbitrário de níveis discretos para poder conduzir o teste. O resultado de tal teste será diferente, dependendo do número de níveis escolhido. Para avaliar a bondade do ajuste às distribuições contínuas, o teste de Kolmogorov-Smirnov é mais apropriado e com maior poder.

Teste da bondade do ajuste a distribuições contínuas: o teste de Kolmogorov-Smirnov

Muitos testes estatísticos requerem que os dados ou os resíduos se ajustem a uma distribuição normal. No Capítulo 9, por exemplo, ilustramos gráficos de diagnóstico para examinar os resíduos de uma regressão linear (Figuras 9.5 e 9.6). Embora tenhamos afirmado que a Figura 9.5A ilustra a distribuição esperada dos resíduos para um modelo linear com distribuição normal dos erros, gostaríamos de ter um teste quantitativo para aquela afirmação. O teste de Kolmogorov-Smirnov é um desses testes.

O teste de Kolmogorov-Smirnov compara a distribuição de uma amostra de dados empíricos com uma distribuição hipotética, como a do tipo normal. Especificamente, esse teste compara as funções de distribuição cumulativa (FDC) empírica com a hipotética. A FDC é definida como a função $F(Y) = P(X < Y)$ para uma variável aleatória X. Em palavras, se X é uma variável aleatória com uma função densidade de probabilidade (FDP) $f(X)$, então a função de distribuição cumulativa $F(Y)$ é igual à área sob $f(X)$ no intervalo $X < Y$ (*ver* Capítulo 2 para uma discussão adicional sobre FDP e FDC). Neste livro, usamos muitas vezes a função de distribuição cumulativa – a probabilidade de cauda, ou valor de P, é a área sob a curva além do ponto Y, que é igual a $1 - F(Y)$.

A Figura 11.4 ilustra duas FDC. A primeira é do tipo empírica para os valores dos resíduos de uma regressão linear, descrita no Capítulo 9, do $\log_{10}$(riqueza de espécies de plantas) pelo $\log_{10}$(área da ilha). A segunda é uma FDC hipotética que seria esperada caso os resíduos tivessem distribuição normal com média = 0 e desvio-padrão = 0,31 (esses valores são estimados a partir dos próprios resíduos). A FDC dos dados empíricos não é uma curva suave, pois é proveniente de uma distribuição normal subjacente com infinitamente muitos pontos.

O teste de Kolmogorov-Smirnov é bicaudal. A hipótese nula é de que a FDC observada $[F_{obs}(Y)]$ = a FDC da distribuição hipotética $[F_{dist}(Y)]$. A estatística do teste é o valor absoluto da diferença vertical máxima entre a FDC observada e a hipotética (seta na Figura 11.4); basicamente, uma mede-se essas distâncias em todos os pontos observados e pega-se a maior delas.

Para a Figura 11.4, a distância máxima é 0,148. Esta diferença máxima é comparada a uma tabela de valores críticos (p. ex., Lilliefors, 1967) para um dado tamanho amostral e valor de P. Se a distância máxima for menor que um valor crítico associado, não podemos rejeitar a hipótese nula de que nossos dados não diferem da distribuição esperada. Em nosso exemplo, o valor crítico para $P = 0,05$ é 0,206. Como a diferença máxima observada foi de apenas 0,148, concluímos que a distribuição dos resíduos da nossa regressão são consistentes com a hipótese de que os erros reais vêm de uma distribuição normal (não rejeitamos a hipótese nula H_0). Por outro lado, se a diferença máxima fosse maior que 0,206, poderíamos rejeitar a hipótese nula e deveríamos questionar se o modelo de regressão linear, que assume que os erros têm distribuição normal, é o apropriado para analisar os dados.

O teste de Kolmogorov-Smirnov não é limitado à distribuição normal, ele pode ser usado para qualquer distribuição contínua. Se você quer testar se seus

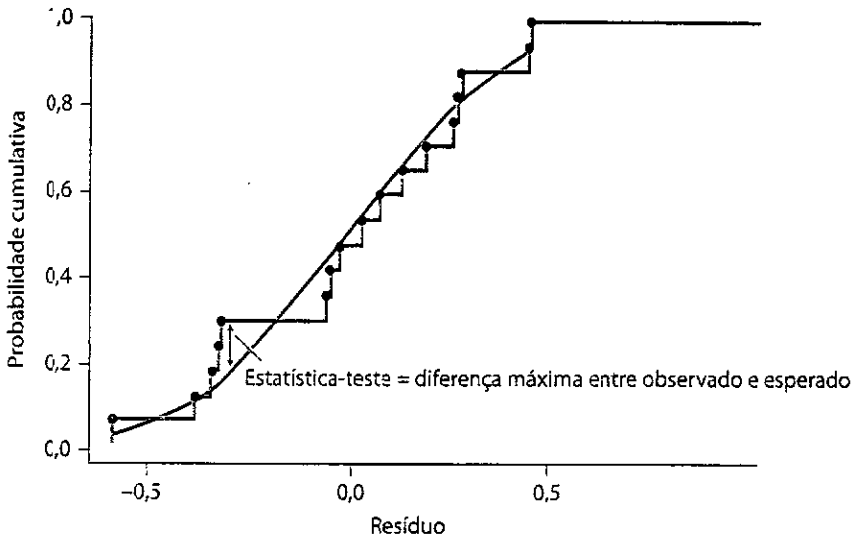


Figura 11.4 Teste de bondade do ajuste para uma distribuição contínua. A comparação da função de distribuição cumulativa dos resíduos observados (símbolos pretos, linha cinza) com a distribuição esperada caso elas viessem de uma distribuição normal (linha verde). Os resíduos são calculados a partir da regressão espécie-área do $\log_{10}$(número de espécies) pelo $\log_{10}$(área das ilhas) de espécies de plantas das ilhas de Galápagos (dados na Tabela 8.2; *ver* a Figura 9.6). O teste da bondade do ajuste de Kolmogorov-Smirnov compara essas duas distribuições, como indicado pela seta de duas pontas. A diferença máxima para esses dados é 0,148, comparado ao valor crítico de 0,206 para um teste a um $P = 0,05$. Como a hipótese nula não pode ser rejeitada, os resíduos parecem seguir uma distribuição normal, que é uma das premissas da regressão linear (Capítulo 9). O teste de Kolmogorov-Smirnov pode ser usado para comparar uma amostra de dados contínuos a qualquer distribuição contínua, como a normal, exponencial ou a log-normal (Capítulo 2).

dados têm distribuição log-normal ou exponencial, por exemplo, pode usar o teste de Kolmogorov-Smirnov para comparar a FDC empírica com uma FDC log-normal ou exponencial (Lilliefors, 1969).

RESUMO

Os dados categóricos de delineamentos tabulares são um resultado comum na pesquisa ecológica e ambiental. Tais dados podem ser analisados com facilidade usando os testes familiares de chi-quadrado e teste-*G*. Os modelos log-lineares hierárquicos ou árvores de classificação podem testar hipóteses detalhadas considerando associações entre variáveis categóricas. Um subconjunto desses métodos também pode ser usado para testar se a distribuição dos dados e dos resíduos é ou não consistente com, ou se são ajustados por, uma distribuição teórica em particular. Alternativas bayesianas a esses testes têm sido desenvolvidas e tais métodos permitem a estimativa de distribuições de frequências e parâmetros esperados em modelos log-lineares.

12

Análise de Dados Multivariados

Todos os métodos analíticos descritos até então se aplicam apenas para uma única variável resposta, ou dados univariados. Vários estudos ecológicos e ambientais, contudo, geram duas ou mais variáveis respostas. Em particular, frequentemente analisamos como múltiplas variáveis respostas – **dados multivariados** – estão relacionadas de maneira simultânea com uma ou mais variáveis preditoras. Por exemplo, uma análise univariada do tamanho da planta-jarro pode basear-se em uma única variável: altura do jarro. Uma análise multivariada seria fundamentada em múltiplas variáveis: altura do jarro, abertura da boca, larguras do tubo e da quilha, comprimento e envergadura da asa (*ver* Tabela 12.1). Como essas variáveis respostas são medidas no mesmo indivíduo, não são independentes entre si. Os métodos estatísticos para analisar dados univariados – regressão, ANOVA, testes chi-quadrado e similares – podem ser inapropriados para analisar multivariados. Neste último capítulo, apresentaremos várias classes de métodos que os ecólogos e cientistas da área ambiental usam para descrever e analisar dados multivariados.

Assim como com os métodos univariados que discutimos nos Capítulos 9 a 11, podemos apenas pincelar e descrever os elementos importantes das formas mais comuns de análise multivariada. Gauch (1982) e Manly (1991) apresentam técnicas clássicas de análise multivariada comumente usadas por ecólogos e cientistas da área ambiental, enquanto Legendre & Legendre (1998) fornecem um tratamento mais aprofundado dos desenvolvimentos até meados da década de 90. O desenvolvimento de novos métodos multivariados, assim como a avaliação das técnicas existentes, é uma área ativa da pesquisa estatística. Ao longo deste capítulo, destacamos alguns desses novos métodos e os debates acerca deles.

ABORDANDO DADOS MULTIVARIADOS

Os dados multivariados se parecem muito com os univariados: consistem de uma ou mais variáveis independentes (preditoras) e duas ou mais variáveis dependentes (respostas). A distinção entre dados univariados e multivariados recai em grande parte sobre como os dados são organizados e analisados, e não em como são coletados.[1]

A maioria dos estudos ecológicos e ambientais produz dados multivariados. Exemplos incluem um conjunto de diferentes medidas morfológicas, alométricas e fisiológicas realizadas em cada uma de várias plantas atribuídas a diferentes grupos de tratamento, uma lista de espécies e suas abundâncias registradas em múltiplas estações de amostragem ao longo de um rio, um conjunto de variáveis ambientais (p. ex., temperatura, umidade, precipitação, etc.) e um conjunto correspondente de abundâncias de organismos ou características medidas em vários locais ao longo de um gradiente geográfico. Em alguns casos, entretanto, o objetivo explícito de uma análise multivariada pode alterar o delineamento amostral. Por exemplo, um estudo de marcação-recaptura de várias espécies em vários locais não seria delineado da mesma forma que um de marcação-recaptura de uma única espécie, pois diferentes métodos de análise poderão ser usados para espécies raras *versus* espécies comuns.

Dados multivariados podem ser quantitativos ou qualitativos, contínuos, ordenados, ou categóricos. Nossa atenção, neste capítulo, será focada em variáveis quantitativas contínuas. As extensões para variáveis multivariadas qualitativas são discutidas por Gifi (1990).

A necessidade da álgebra de matrizes

A matemática das análises multivariadas é melhor expressa por meio de **álgebra de matrizes**, a qual nos permite empregar equações semelhantes às usadas anteriormente nas análises univariadas. A similaridade é mais do que superficial, as interpretações dessas equações são próximas, e muitas das equações usadas para métodos estatísticos univariados também podem ser escritas com notação de matrizes. Resumimos o básico da notação e da álgebra de matrizes no Apêndice.

Uma variável aleatória multivariada **Y** é um conjunto de n variáveis univariadas $\{Y_1, Y_2, Y_3, .., Y_n\}$ que são todas medidas para a mesma unidade observacional ou experimental, como uma única ilha, planta, ou estação amostral.

[1] Na verdade, você já foi apresentado a dados multivariados nos Capítulos 6, 7 e 10. Em um delineamento de medidas repetidas, múltiplas observações são obtidas do mesmo indivíduo em momentos diferentes. Em um delineamento em blocos randomizados ou em parcelas subdivididas, determinados tratamentos são agrupados fisicamente no mesmo bloco. Como as variáveis respostas medidas dentro de um bloco ou em um único indivíduo não são independentes entre si, a análise de variância tem de ser modificada para levar em conta essa estrutura nos dados (*ver* Capítulo 10). Neste capítulo, descreveremos alguns métodos multivariados nos quais não há uma análise univariada correspondente.

Indicamos uma variável aleatória multivariada com uma letra maiúscula e em negrito, enquanto continuamos a escrever variáveis aleatórias univariadas com letras maiúsculas em itálico. Cada observação em **Y** é escrita como $\mathbf{y}_i$, que é um **vetor linha.** Os elementos $y_{i,j}$ desse vetor linha correspondem à medida da j-ésima variável Y_j (com j variando de 1 até n variáveis) medida para o indivíduo i (com i variando de 1 até m observações individuais ou unidades experimentais):

$$\mathbf{y}_i = [y_{i,1}, y_{i,2}, y_{i,3}, \cdots y_{i,j}, \ldots, y_{i,n}] \tag{12.1}$$

Por conveniência, em geral omitimos o subscrito i na Equação 12.1. O vetor linha **y** é um exemplo de **matriz** com uma linha e n colunas. Cada observação $y_{i,j}$ dentro dos colchetes é chamada **elemento** da matriz (ou vetor linha) **y**.

Um ecólogo organizaria dados multivariados típicos em uma planilha simples com m linhas, cada uma representando um indivíduo diferente, e n colunas, cada uma representando uma diferente variável medida. Cada linha da matriz corresponde ao vetor linha $\mathbf{y}_i$ e toda a tabela é a matriz **Y**. Um exemplo de conjunto de dados multivariados é apresentado na Tabela 12.1. Cada uma das $m = 89$ linhas representa uma planta diferente, e cada uma das $n = 10$ colunas, uma variável morfológica diferente da que foi medida em cada planta.

Várias matrizes serão recorrentes ao longo desse capítulo. A Equação 12.1 é o vetor básico n-dimensional descrevendo uma única observação: $\mathbf{y} = [y_1, y_2, y_3, \ldots, y_n]$. Também definimos um **vetor de médias** como:

$$\overline{\mathbf{Y}} = [\overline{Y}_1, \overline{Y}_2, \overline{Y}_3, \ldots, \overline{Y}_n] \tag{12.2}$$

TABELA 12.1 Dados multivariados para a planta carnívora *Darlingtonia californica*

Local	Planta	Altura	Boca	Tubo	Quilha	Asa1	Asa2	Envergadura	Massa do capuz	Massa do tubo	Massa da asa
TJH	1	654	38	17	6	85	76	55	1,38	3,54	0,29
TJH	2	413	22	17	6	55	26	60	0,49	1,48	0,06
A	A	A	A	A	A	A	A	A	A	A	A

Cada linha representa uma única observação de uma planta lírio-cobra (*Darlingtonia californica*) medida em um local particular (charco T.J. Howell, TJH) nas Montanhas Siskiyou no sudeste de Oregon. Cada coluna representa uma variável morfológica diferente. As medidas de cada planta incluíram sete variáveis morfológicas (todas em mm): altura (altura), abertura da boca (boca), diâmetro do tubo (tubo), diâmetro da quilha (quilha), comprimento de cada uma das duas "asas" (asa1, asa2) que formam o apêndice com forma de cauda de peixe, a distância entre a ponta das duas asas (envergadura). As plantas foram secas e o capuz (massa do capuz), o tubo (massa do tubo) e o apêndice cauda de peixe (massa da asa) foram pesados (± 0,01g). O conjunto de dados total, coletado em 2000 por A. Ellison, R. Emerson, e H. Steinhoff, consiste de 89 plantas medidas em 4 locais. Esse é um típico conjunto de dados multivariado: as diferentes variáveis medidas em um mesmo indivíduo não são independentes umas das outras.

que é o vetor n-dimensional de médias amostrais (calculadas como descrito no Capítulo 3) de cada Y_j. Os elementos desse vetor linha são as médias de cada uma das n variáveis, Y_j. Em outras palavras, $\bar{\mathbf{Y}}$ é igual ao conjunto de médias das amostras de cada coluna da nossa matriz de dados original **Y**. Em uma análise univariada, a medida correspondente é a média da amostra de uma única variável resposta, que é uma estimativa da média real μ da população. De forma similar, o vetor $\bar{\mathbf{Y}}$ é uma estimativa do vetor $\mu = [\mu_1, \mu_2, \mu_3, ..., \mu_n]$ das médias reais da população para cada uma das n variáveis.

Precisamos de três matrizes para descrever a variação entre as observações $y_{i,j}$. A primeira é a **matriz de variância e covariância**, C (*ver* Nota de rodapé 5, Capítulo 9). Trata-se de uma **matriz quadrada** cujos elementos da diagonal principal são as variâncias amostrais de cada variável (calculada da forma usual; *ver* Capítulo 3, Equação 3.9). Os elementos fora da diagonal são as covariâncias amostrais entre todos os possíveis pares de variáveis (calcule como fizemos para a regressão e ANOVA: *ver* Capítulo 9, Equação 9.10).

$$\mathbf{C} = \begin{bmatrix} s_1^2 & c_{1,2} & \cdots & c_{1,n} \\ c_{2,1} & s_2^2 & \cdots & c_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & c_{n,2} & \cdots & s_n^2 \end{bmatrix} \tag{12.3}$$

Na Equação 12.3, s_j^2 é a variância amostral da variável Y_j e $c_{j,k}$ é a covariância amostral entre variáveis Y_j e Y_k. Note que a covariância amostral entre as variáveis Y_j e Y_k é idêntico à covariância amostral entre as variáveis Y_k e Y_j. Portanto, os elementos da matriz de variância e covariância são **simétricos**, e são imagens-espelho uns dos outros no outro lado da diagonal ($c_{j,k} = c_{k,j}$).

A segunda matriz que usamos para descrever as variâncias é a dos desvios-padrões amostrais:

$$\mathbf{D(s)} = \begin{bmatrix} \sqrt{s_1^2} & 0 & \cdots & 0 \\ 0 & \sqrt{s_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{s_n^2} \end{bmatrix} \tag{12.4}$$

D(s) é uma **matriz diagonal** cujos elementos são todos 0, exceto aqueles na diagonal principal. Os elementos "não zero" são os desvios-padrões amostrais de cada Y_j.

Por último, precisamos de uma matriz de correlações amostrais:

$$\mathbf{P} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1,n} \\ r_{2,1} & 1 & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & 1 \end{bmatrix} \quad (12.5)$$

Nesta, cada elemento $r_{j,k}$ é a correlação amostral (*ver* Capítulo 9, Equação 9.19) entre as variáveis Y_j e Y_k. Como a correlação de uma variável com ela mesma é = 1, os elementos da diagonal de **P** são iguais a 1. Assim como a de variância e covariância, a matriz de correlações é simétrica ($r_{j,k} = r_{k,j}$).

COMPARANDO MÉDIAS MULTIVARIADAS

Comparando médias multivariadas de duas amostras: Teste T^2 de Hotelling

Apresentamos os conceitos de análise multivariada através de uma simples extensão do teste-*t* univariado para o caso multivariado. O teste-*t* clássico é usado para testar a hipótese nula de que o valor médio de uma única variável não difere entre dois grupos ou populações. Por exemplo, poderíamos testar a simples hipótese de que a altura do jarro da *Darlingtonia californica* (Tabela 12.1) não difere entre populações amostradas em Day's Gulch (DG) e no charco T.J.Howell (TJH). Em cada local, 25 plantas foram amostradas de forma aleatória, das quais primeiro calculamos a altura média (DG = 618,8 mm; TJH = 610,4 mm) e o desvio-padrão (DG = 100,6 mm; TJH = 83,7 mm). Com base no teste-*t* padrão, as plantas dessas duas populações não diferem significativamente quanto à altura (t = 0,34, com 48 graus de liberdade, P = 0,74). Poderíamos fazer testes-*t* adicionais para cada uma das variáveis na Tabela 12.1 com ou sem correções para testes múltiplos (*ver* Capítulo 10). Entretanto, estamos mais interessados em perguntar se a morfologia geral das plantas – quantificada como um vetor de médias de todas as variáveis morfológicas tomadas em conjunto – difere entre os dois locais. Em outras palavras, desejamos testar a hipótese nula de que os vetores das médias dos dois grupos são iguais.

O teste T^2 de Hotelling (Hotelling, 1931) é uma generalização do teste-*t* univariado para dados multivariados. Primeiramente, calculamos as médias de cada grupo para cada uma das variáveis de interesse Y_j e as juntamos, então, em dois **vetores coluna** de médias (*ver* Equação 12.2), $\overline{\mathbf{Y}}_1$ (para TJH) e $\overline{\mathbf{Y}}_2$ (para DG). Note que um vetor-coluna é simplesmente um vetor-linha que foi **transposto** – as colunas e linhas são intercambiadas. As regras da álgebra de matrizes, explicadas no Apêndice, às vezes requerem que os vetores sejam colunas únicas, ou por vezes requerem que os vetores sejam linhas únicas. Mas os dados

são os mesmos, independente de quando são representados como vetores-linha ou vetores-coluna.

Para este exemplo, temos sete variáveis morfológicas: altura do jarro, abertura da boca, larguras do tubo e da quilha, comprimento e envergadura das asas do apêndice cauda de peixe na boca do jarro (Tabela 12.1), cada uma medida em dois locais.

$$\bar{Y}_1 = \begin{bmatrix} 610{,}0 \\ 31{,}2 \\ 19{,}9 \\ 6{,}7 \\ 61{,}5 \\ 59{,}9 \\ 77{,}9 \end{bmatrix} \qquad \bar{Y}_2 = \begin{bmatrix} 618{,}8 \\ 33{,}1 \\ 17{,}9 \\ 5{,}6 \\ 82{,}4 \\ 79{,}4 \\ 84{,}2 \end{bmatrix} \tag{12.6}$$

Por exemplo, a média da altura do jarro em TJH foi 610,0 mm (o primeiro elemento de $\bar{Y}_1$), enquanto a média da abertura da boca do jarro em DG foi 33,1 mm (o segundo elemento de $\bar{Y}_2$). A propósito, as variáveis usadas em uma análise multivariada não têm que ser necessariamente medidas nas mesmas unidades, embora muitas análises multivariadas redimensionam as variáveis em unidades padronizadas, como descrito mais adiante, neste capítulo.

Agora, precisamos de alguma medida de variação. Para o teste-*t*, usamos as variâncias das amostras (s^2) de cada grupo. Para o teste T^2 de Hotelling, empregamos a matriz C de variância e covariância amostral (Equação 12.3) para cada grupo. A matriz de variância e covariância das amostras para o local TJH é:

$$C_1 = \begin{bmatrix} 7011{,}5 & 284{,}5 & 32{,}3 & -12{,}5 & 137{,}4 & 691{,}7 & 76{,}5 \\ 284{,}5 & 31{,}8 & -0{,}7 & -1{,}9 & 55{,}8 & 77{,}7 & 66{,}3 \\ 32{,}3 & -0{,}7 & 5{,}9 & 0{,}6 & -19{,}9 & -6{,}9 & -13{,}9 \\ -12{,}5 & -1{,}9 & 0{,}6 & 1{,}1 & -0{,}9 & -3{,}0 & -5{,}6 \\ 137{,}4 & 55{,}8 & -19{,}9 & -0{,}9 & 356{,}7 & 305{,}6 & 397{,}4 \\ 691{,}7 & 77{,}7 & -6{,}9 & -3{,}0 & 305{,}6 & 482{,}1 & 511{,}6 \\ 76{,}5 & 66{,}3 & -13{,}9 & -5{,}6 & 397{,}4 & 511{,}6 & 973{,}3 \end{bmatrix} \tag{12.7}$$

Essa matriz resume todas as variâncias e covariâncias amostrais das variáveis. Por exemplo, em TJH, a variância amostral da altura do jarro foi 7011,5 mm^2, enquanto a covariância amostral entre altura e abertura da boca do jarro foi 284,5 mm^2. Embora uma matriz de variância e covariância amostral possa ser cons-

truída para cada grupo ($\mathbf{C}_1$ e $\mathbf{C}_2$), o teste T^2 de Hotelling assume que essas duas matrizes são aproximadamente iguais e usa a matriz $\mathbf{C}_p$, que é a estimativa combinada da covariância dos dois grupos:

$$\mathbf{C}_P = \frac{[(m_1 - 1)\mathbf{C}_1 + (m_2 - 1)\mathbf{C}_2]}{(m_1 + m_2 - 2)} \qquad (12.8)$$

onde m_1 e m_2 são os tamanhos amostrais de cada grupo (aqui, $m_1 = m_2 = 25$).

A estatística T^2 de Hotelling é calculada como:

$$T^2 = \frac{m_1 m_2 (\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)^{\mathrm{T}} \mathbf{C}_P^{-1} (\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)}{(m_1 + m_2)} \qquad (12.9)$$

onde $\overline{\mathbf{Y}}_1$ e $\overline{\mathbf{Y}}_2$ são os vetores de médias (Equação 12.2), $\mathbf{C}_p$ é a matriz combinada de variância-covariância amostral (Equação 12.8), o sobrescrito T denota a transposição da matriz, e o sobrescrito (−1) denota a inversão da matriz (*ver* o Apêndice). Note que nesta equação existem dois tipos de multiplicação. O primeiro é a multiplicação de matrizes dos vetores de diferenças entre médias e do inverso da matriz de covariância:

$$(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)^{\mathrm{T}} \mathbf{C}_P^{-1} (\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2)$$

O segundo é a multiplicação escalar desse produto pelo dos tamanhos amostrais $m_1 m_2$.

Seguindo as regras de multiplicação de matrizes, a Equação 12.9 dá o T^2 como um escalar, ou um único número. Uma transformação linear do T^2 produz a familiar estatística teste F:

$$\mathrm{F} = \frac{(m_1 + m_2 - n - 1)T^2}{(m_1 + m_2 - 2)n} \qquad (12.10)$$

onde n é o número de variáveis. Sob a hipótese nula de que os vetores de médias das populações de cada grupo são iguais (i. e., $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$), o F está distribuído como uma variável aleatória F, com n graus de liberdade no numerador e $(m_1 + m_2 - n - 1)$ para o denominador. Então, o teste de hipótese procede da maneira usual (*ver* Capítulo 5). Para as sete variáveis morfológicas, o T^2 é igual a 84,62, com 7 graus de liberdade no numerador e 42 no denominador. A aplicação da Equação 12.10 dá um valor de F de 10,58. O valor crítico de F com 7 graus de liberdade no numerador e 42 no denominador é igual a 2,23, o que é muito menor que o valor de 10,58 observado. Portanto, podemos rejeitar a hipótese nula (com um valor de P de $1{,}3 \times 10^{-7}$) de que os dois vetores de médias das populações, $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$, são iguais.

Comparando médias multivariadas de mais de duas amostras: uma MANOVA simples

Se desejarmos comparar médias univariadas para mais de duas amostras de grupos de tratamento, usamos ANOVA em vez do teste-*t* padrão. Similarmente, se estamos comparando médias multivariadas para mais de dois grupos, usamos uma **ANOVA multivariada**, ou **MANOVA**, em vez do teste T^2 de Hotelling. Como no teste T^2 de Hotelling, temos $j = 1,...,$ até n variáveis ou medidas tomadas de cada indivíduo e $i = 1,...,$ até um total de m total de observações. Na MANOVA, contudo, temos $k = 1$ até g grupos, e $l = 1,...,$ até q observações dentro de cada grupo. As observações multivariadas (Equação 12.1) são denotadas como $\mathbf{y}_{k,l}$, o l-ésimo vetor de observações dentro do k-ésimo grupo. Se o nosso delineamento é balanceado, de modo que exista o mesmo número q de observações em cada um dos g grupos, então o nosso tamanho amostral total $m = gq$.

Por analogia com a ANOVA de um fator (Capítulo 10, Equação 10.6), analisamos o modelo:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{A}_k + \mathbf{e}_{kl} \qquad (12.11)$$

onde $\mathbf{Y}$ é a matriz de medidas (com m linhas de observações e n colunas de variáveis), $\boldsymbol{\mu}$ é a média (geral) da população, $\mathbf{A}_k$ é a matriz de desvios do k-ésimo tratamento para a média geral das amostras e $\mathbf{e}_{kl}$ é o termo do erro – a diferença da média do k-ésimo tratamento do l-ésimo indivíduo daquele grupo de tratamento. Assume-se que o $\mathbf{e}_{kl}$ provém de uma distribuição multivariada normal (*ver* próxima seção). O procedimento é basicamente o mesmo da ANOVA de um fator (Capítulo 10), exceto que, em vez de comparar as médias dos grupos, comparamos os **centroides** dos grupos – as médias multivariadas.[2] A hipótese nula é que as médias dos grupos de tratamento são todas iguais: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = ... = \boldsymbol{\mu}_g$. Como na ANOVA, a estatística do teste é a razão da soma dos quadrados entre grupos, dividida pela soma dos quadrados dentro dos grupos, os quais são calculados diferentemente na MANOVA, como descrito abaixo. Essa razão tem uma distribuição-F (*ver* Capítulo 10).

[2] Em um espaço unidimensional, duas médias podem ser comparadas como a diferença aritmética entre elas, que é o que uma ANOVA faz para dados univariados. Em um espaço bidimensional, o vetor médio de cada grupo pode ser plotado como um ponto $(\bar{X}, \bar{Y})$ em um gráfico cartesiano. Esses pontos representam os centroides (em geral pensados como o centro de gravidade de uma nuvem de pontos) e podemos calcular a distância geométrica entre eles. Por fim, em um espaço com três (ou mais) dimensões, o centroide pode novamente ser localizado por um conjunto de coordenadas cartesianas com uma coordenada para cada dimensão no espaço. A MANOVA compara as distâncias entre esses centroides e testa a hipótese nula de que as distâncias entre os centroides dos grupos não são mais diferentes das esperadas devido à amostragem aleatória.

AS MATRIZES SQPC Em uma ANOVA, as somas dos quadrados são simples números; em uma MANOVA, as somas dos quadrados são matrizes – chamadas de matrizes de **somas dos quadrados e produtos cruzados** (**SQPC**). São matrizes quadradas (número igual de linhas e colunas) cujos elementos da diagonal correspondem às somas dos quadrados para cada variável e os elementos fora da diagonal, às somas dos produtos cruzados para cada par de variáveis. As matrizes SQPC são diretamente análogas às de variância-covariância amostrais (Equação 12.3), porém, precisamos de três delas.

A matriz SQPC entre grupos é a **H**:

$$\mathbf{H} = q\sum_{k=1}^{g}(\overline{\mathbf{Y}}_{k.} - \overline{\mathbf{Y}}_{..})(\overline{\mathbf{Y}}_{k.} - \overline{\mathbf{Y}}_{..})^{\mathrm{T}} \qquad (12.12)$$

Na Equação 12.12, o $\overline{\mathbf{Y}}_{k.}$ é o vetor de médias das amostras no grupo k das q observações naquele grupo:

$$\overline{\mathbf{Y}}_{k.} = \frac{1}{q}\sum_{l=1}^{q}\mathbf{y}_{kl}$$

e o $\overline{\mathbf{Y}}_{..}$ é o vetor de médias das amostras de todos os grupos-tratamento:

$$\overline{\mathbf{Y}}_{..} = \frac{1}{kq}\sum_{k=1}^{g}\sum_{l=1}^{q}\mathbf{y}_{kl}$$

A matriz SQPC dentro de grupos é a matriz **E**:

$$\mathbf{E} = \sum_{k=1}^{g}\sum_{l=1}^{q}(\mathbf{y}_{kl} - \overline{\mathbf{Y}}_{k.})(\mathbf{y}_{kl} - \overline{\mathbf{Y}}_{k.})^{\mathrm{T}} \qquad (12.13)$$

com $\overline{\mathbf{Y}}_{k.}$ novamente sendo o vetor de médias das amostras do grupo-tratamento k. Em uma análise univariada, o análogo do **H** na Equação 12.12 é a soma dos quadrados entre grupos (Equação 10.2, Capítulo 10), e o análogo de **E** na Equação 12.13 é a soma dos quadrados dentro de grupos (Equação 10.3, Capítulo 10). Por fim, calculamos a matriz SQPC total:

$$\mathbf{T} = \sum_{k=1}^{g}\sum_{l=1}^{q}(\mathbf{y}_{kl} - \overline{\mathbf{Y}}_{..})(\mathbf{y}_{kl} - \overline{\mathbf{Y}}_{..})^{\mathrm{T}} \qquad (12.14)$$

ESTATÍSTICAS DO TESTE Quatro estatísticas-teste são geradas a partir das matrizes **H** e **E**: lambda de Wilk, traço de Pillai, traço de Hotelling-Lawley e maior raiz de Roy (*ver* Scheiner, 2001). Todas são calculadas como a seguir:

$$\text{Lambda de Wilk} = \Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}$$

onde | | denota o determinante de uma matriz (*ver* Apêndice).

$$\text{Traço de Pillai} = \sum_{i=1}^{s}\left(\frac{\lambda_i}{\lambda_i+1}\right) = \mathit{traço}\left[(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}\right]$$

onde s é o menor valor entre qualquer um dos dois, ou os graus de liberdade (número de grupos $g-1$) ou o número de variáveis n, λ_i é o i-ésimo autovalor de $\mathbf{E}^{-1}\mathbf{H}$, e *traço* é o traço de uma matriz (*ver* Apêndice).

$$\text{Traço de Hotelling-Lawley} = \sum_{i=1}^{n}\lambda_i = \mathit{traço}\left(\mathbf{E}^{-1}\mathbf{H}\right)$$

e

$$\text{Maior raiz de Roy} = \theta = \frac{\lambda_1}{\lambda_1+1}$$

onde λ_1 é o maior (primeiro) autovalor de $\mathbf{E}^{-1}\mathbf{H}$.

Para tamanhos amostrais grandes, o lambda de Wilk, o traço de Hotelling-Lawley e o traço de Pillai convergem para o mesmo valor de P, embora o traço de Pillai seja o mais clemente às violações de pressupostos, como a normalidade multivariada. A maioria dos programas irá dar todas essas estatísticas-teste.

Cada uma dessas quatro estatísticas-teste pode ser transformada em uma estatística-F e testada como na ANOVA. Contudo, seus graus de liberdade não são os mesmos. O lambda de Wilk, o traço de Pillai e o traço de Hotelling-Lawley são calculados a partir de todos os autovalores e, então, têm um tamanho amostral de nm (número de variáveis n × número de amostras m) com $n(g-1)$ graus de liberdade no numerador (g é o número de grupos). A maior raiz de Roy usa apenas o primeiro autovalor, então seus graus de liberdade no numerador são apenas n. Os graus de liberdade para o lambda de Wilk com frequência são fracionários, e a estatística-F associada é apenas uma aproximação (Harris, 1985). A escolha entre essas estatísticas de teste da MANOVA não é crucial. Elas em geral dão resultados muito similares, como na análise dos dados da *Darlingtonia* (Tabela 12.2).

COMPARAÇÕES ENTRE GRUPOS Assim como na ANOVA, se a MANOVA dá resultados significativos, você pode estar interessado em determinar quais grupos em particular diferem uns dos outros. Para comparações *a posteriori*, o teste T^2 de Hotelling com o valor crítico ajustado usando a correção de Bonferroni (*ver* Capítulo 10) pode ser usado para cada comparação par-a-par. A análise discriminante (descrita posteriormente neste Capítulo) também pode ser usada para determinar quão bem os grupos podem ser separados.

TABELA 12.2 Resultados de uma MANOVA de um fator

I. Matriz H

	Altura	Boca	Tubo	Quilha	Asa1	Asa2	Envergadura
Altura	35.118,38	5.584,16	−1.219,21	−2.111,58	17.739,19	11.322,59	18.502,78
Boca	5.584,16	1.161,14	−406,04	−395,59	2.756,84	1.483,22	1.172,86
Tubo	−1.219,21	−406,04	229,25	127,10	−842,68	−432,92	659,62
Quilha	−2.111,58	−395,59	127,10	142,49	−1.144,14	−706,07	−748,92
Asa1	17.739,19	2.756,84	−842,68	−1.144,14	12.426,23	9.365,82	10.659,12
Asa2	11.322,59	1.483,22	−432,92	−706,07	9.365,82	7.716,96	8.950,57
Envergadura	18.502,78	1.172,86	659,62	−748,92	10.659,12	8.950,57	21.440,44

II. Matriz E

	Altura	Boca	Tubo	Quilha	Asa1	Asa2	Envergadura
Altura	834.836,33	27.526,88	8.071,43	1.617,17	37.125,87	46.657,82	18.360,69
Boca	27.526,88	2.200,17	324,05	−21,91	4.255,27	4.102,43	3.635,34
Tubo	8.071,43	324,05	671,82	196,12	305,16	486,06	375,09
Quilha	1.617,17	−21,91	196,12	265,49	−219,06	−417,36	−632,27
Asa1	37.125,87	4.255,27	305,16	−219,06	31.605, 96	28.737,10	33.487,39
Asa2	46.657,82	4.102,43	486,06	−417,36	28.737,10	39.064,26	41.713,77
Envergadura	18.360,69	3.635,34	375,09	−632,27	33.487,39	41.713,77	86.181,61

III. Estatísticas-teste

Estatística	Valor	F	gl no numerador	gl no denominador	Valor de *P*
Traço de Pillai	1,11	6,45	21	231	3×10^{-14}
Lambda de Wilk	0,23	6,95	21	215,91	3×10^{-15}
Traço de Hotelling-Lawley	2,09	7,34	21	221	3×10^{-16}
Maior autovalor de Roy	1,33	14,65	7	77	5×10^{-12}

Os dados originais são medidas de 7 variáveis morfológicas em 89 indivíduos do lírio-cobra, *Darlingtonia californica* coletadas em 4 locais (Tabela 12.1). H é a matriz (SQPC) de somas dos quadrados entre grupos e dos produtos cruzados. Os elementos da diagonal são as somas dos quadrados de cada variável, os elementos externos à diagonal são as somas dos produtos cruzados para cada par de variáveis (Equação 12.12). A matriz H é análoga ao cálculo univariado da soma dos quadrados entre grupos (Equação 10.2). A matriz E é a de SQPC dentro de grupos. Os elementos da diagonal são os desvios internos de grupos de cada variável, os elementos externos à diagonal são os resíduos dos produtos cruzados para cada par de variáveis (Equação 12.13). A matriz E é análoga ao cálculo univariado da soma dos quadrados dos resíduos (Equação 10.3). Nessas matrizes, partimos da boa prática de reportar apenas dígitos significativos a fim de evitar erros de arredondamento em nossos cálculos. Se fossemos reportar esses dados em uma publicação científica, eles não teriam mais dígitos significativos do que tivemos em nossas medidas (*ver* Tabela 12.1). As quatro estatísticas-teste (traço de Pillai, lambda de Wilk, traço de Hotelling-Lawley e maior raiz de Roy) representam maneiras diferentes de testar as diferenças entre os grupos em uma análise de variância multivariada (MANOVA). Todas essas medidas geraram valores de *P* muito pequenos, sugerindo que os vetores de 7 elementos das medidas morfológicas da *Darlingtonia* diferem significativamente entre os locais.

Todos os delineamentos de ANOVA descritos no Capítulo 10 têm análogos diretos na MANOVA, quando as variáveis respostas são multivariadas. A única diferença é que as matrizes SQPC são usadas em vez das somas dos quadrados entre e dentro de grupos. Pode ser difícil calcular matrizes SQPC para delineamentos experimentais complexos (Harris, 1985; Hand & Taylor, 1987). Scheiner (2001) resume técnicas de MANOVA para ecólogos e cientistas da área ambiental.

PRESSUPOSTOS DA MANOVA Além dos pressupostos usuais da ANOVA (observações são amostradas independente e aleatoriamente, os erros dentro dos grupos são iguais entre eles e têm distribuição normal), a MANOVA tem dois pressupostos adicionais. Primeiro, semelhante aos requerimentos do teste T^2 de Hotelling, as covariâncias são iguais entre os grupos (o pressuposto da **esfericidade**). Segundo, as variáveis multivariadas usadas na análise e os termos do erro $\mathbf{e}_H$ na Equação 12.11 devem se conformar a uma distribuição normal multivariada. Na próxima seção, descrevemos a distribuição normal multivariada e como testar os desvios dessa distribuição.

Como o traço de Pillai é robusto a violações modestas desses pressupostos, na prática, a MANOVA é válida se os dados não desviam substancialmente da esfericidade ou da normalidade multivariada. A Análise de Similaridade (ANOSIM) é um método alternativo à MANOVA (Clarke & Green, 1988; Clarke, 1993), que não depende da normalidade multivariada, mas pode ser usado apenas para delineamentos de um fator e totalmente cruzados, ou em delineamentos com dois fatores aninhados. A ANOSIM tem menos poder (alta probabilidade de erro estatístico do Tipo II) na presença de gradientes fortes nos dados (Somerfield et al., 2002).

A DISTRIBUIÇÃO NORMAL MULTIVARIADA

A maioria dos métodos multivariados para teste de hipóteses requer que os dados analisados se ajustem a uma **distribuição normal multivariada** (ou **multinormal**). A distribuição normal multivariada é análoga da distribuição normal (ou gaussiana) para a variável multidimensional $\mathbf{Y} = [Y_1, Y_2, Y_3, \ldots, Y_n]$. Como vimos no Capítulo 2, a distribuição normal tem dois parâmetros: μ (a média) e σ^2 (a variância). Em contraste, a distribuição normal multivariada[3] é

(Continua)

[3] A função de densidade de probabilidade para uma distribuição normal multivariada assume um vetor aleatorio n-dimensional $\mathbf{Y}$ de variáveis aleatórias $[Y_1, Y_2, \ldots, Y_n]$ com vetor de médias $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_n]$ e matriz de variância e covariância

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \gamma_{12} & \cdots & \gamma_{1n} \\ \gamma_{21} & \sigma_2^2 & \cdots & \gamma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

definida pelo vetor de médias $\boldsymbol{\mu}$ e pela matriz de covariância $\boldsymbol{\Sigma}$. O número de parâmetros na distribuição normal multivariada depende do de variáveis aleatórias Y_i em **Y**. Como se sabe, uma distribuição normal unidimensional tem uma curva em forma de sino (*ver* Figura 2.6). Para uma variável bidimensional $\mathbf{Y} = [Y_1, Y_2]$, a distribuição normal bivariada parece com um sino ou chapéu (Figura 12.1). A maioria da massa de probabilidade está concentrada perto do centro do chapéu, o mais próximo das médias das duas variáveis. Conforme movemos em qualquer direção para a aba do chapéu, a densidade de probabilidade se torna cada vez mais fina. Embora seja difícil desenhar uma distribuição multivariada para variáveis de mais de duas dimensões, podemos calcular e interpretá-las da mesma maneira.

Para a maioria dos objetivos, é conveniente usar a distribuição normal multivariada padronizada, que tem um vetor de médias $\boldsymbol{\mu} = [0]$. A distribuição normal bivariada ilustrada na Figura 12.1 é, na verdade, uma distribuição normal bivariada padronizada. Também podemos ilustrar essa distribuição como um gráfico de contorno (Figura 12.1E) de elipses concêntricas. A elipse central corresponde a uma fatia do pico da distribuição e, conforme descemos no chapéu, as fatias elípticas se tornam mais largas. Quando todas as variáveis Y_n em **Y** são independentes e não correlacionadas, todos os coeficientes de correlação fora da diagonal r_{mn} na matriz de correlação **P** são = 0, e as fatias elípticas são perfeitamente circulares. Contudo, as variáveis na maioria dos conjuntos de dados multivariados são correlacionadas. Portanto, as fatias em geral têm forma elíptica, como na Figura 12.1E. Quanto mais forte a correlação entre um par de variáveis, mais prolongados são os contornos da elipse. A ideia pode ser estendida para mais de duas variáveis em **Y** e, neste caso, as fatias seriam elipsoides ou hiperelipsoides no espaço multidimensional.

Teste de normalidade multivariada

Vários testes multivariados assumem que os dados multivariados ou que seus resíduos tenham distribuição normal multivariada. Os testes para normalidade univariada (descritos no Capítulo 11) são usados rotineiramente com regressão,

---

[3] (*Continuação*)

na qual σ_i^2 é a variância de Y_i e γ_{ij} é a covariância de Y_i e Y_j. Para este vetor **Y**, a função densidade de probabilidade para a distribuição normal multivariada é

$$f(\mathbf{Y}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} e^{\left[-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right]}$$

Ver o Apêndice para detalhes sobre determinantes ($|\ |$), inversos (matrizes elevadas à potência (−1)) e transposição (T) de matrizes. O número de parâmetros necessários para definir uma distribuição normal multivariada é $2n + n(n-1)/2$, onde n é o número de variáveis em **Y**. Note como essa equação se parece com a função de densidade de probabilidade para a distribuição normal univariada (Nota de Rodapé 10, Capítulo 2, e Tabela 2.4).

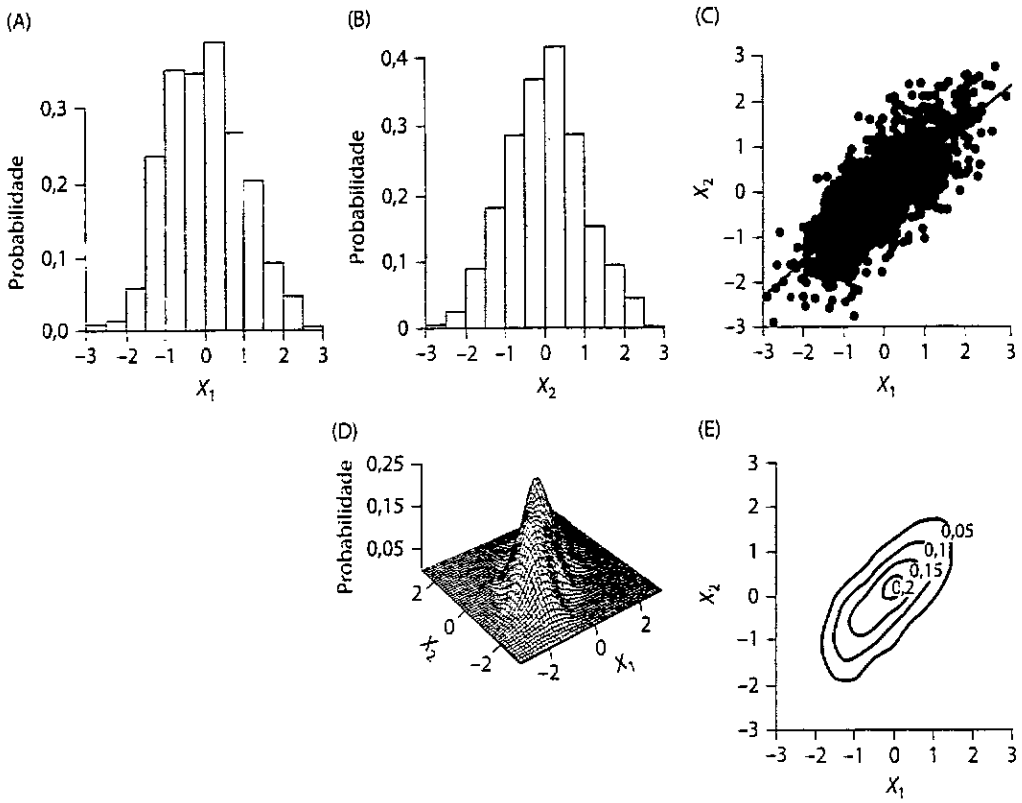


Figura 12.1 Exemplo de uma distribuição normal multivariada. A variável aleatória multivariada **X** é um vetor de duas variáveis univariadas, X_1 e X_2, cada uma com 5.000 observações: $\mathbf{X} = [X_1, X_2]$. O vetor de médias $\boldsymbol{\mu} = [0, 0]$, pois ambos X_1 e X_2 têm médias = 0. O desvio-padrão tanto de X_1 quanto de X_2 é = 1, uma vez que a matriz **D**(**s**) dos desvios-padrões amostrais $= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. (A, B) Esses histogramas ilustram as distribuições de frequência de X_1 (A) e X_2 (B). A correlação entre X_1 e $X_2 = 0{,}75$ e é ilustrada no gráfico de dispersão (C). Assim, a matriz **P** das correlações amostrais $= \begin{bmatrix} 1 & 0{,}75 \\ 0{,}75 & 1 \end{bmatrix}$. A distribuição de probabilidade multivariada conjunta de **X** é apresentada como um gráfico de rede em (D). Essa distribuição também pode ser representada como um gráfico de contorno (E), onde cada contorno representa uma fatia através do gráfico de rede. Assim, para X_1 e X_2 próximos de 0, a marca do contorno = 0,2 significa que a probabilidade conjunta de se obter esses dois valores é aproximadamente 0,2. A direção e a excentricidade das elipses (ou fatias) do contorno refletem a correlação entre as variáveis X_1 e X_2 (painel C). Se X_1 e X_2 fossem completamente não correlacionados, os contornos seriam circulares. Se X_1 e X_2 fossem correlacionados, os contornos iriam colapsar em uma única linha reta no espaço bivariado.

ANOVA e outras análises estatísticas univariadas, mas testes para normalidade multivariada raras vezes são realizados. Isso não é pela falta de disponibilidade desses testes. De fato, mais de cinquenta desses testes de normalidade multivariada foram propostos (revisão em Koizol, 1986; Gnanadesikan, 1997; e Mecklin & Mundfrom, 2003). No entanto, esses testes não estão incluídos nos pacotes de programas estatísticos, e até hoje não foi encontrado nenhum que leve em conta o grande número de formas com que os dados multivariados podem desviar da normalidade (Mecklin & Mundfrom, 2003). Como esses testes são complexos, e geralmente dão resultados conflitantes, eles têm sido chamados de "curiosidades acadêmicas, pouco usados por estatísticos praticantes" (Horswell, 1990, citado em Mecklin & Mundfrom, 2003).

Um atalho comum e simples para testar a normalidade multivariada é verificar se cada variável individual dentro do conjunto de dados multivariado tem distribuição normal (*ver* o Capítulo 11 para detalhes). Se qualquer uma das variáveis individuais não tiver distribuição normal, então não é possível que o conjunto de dados multivariado tenha uma distribuição normal multivariada. Contudo, o contrário não é verdadeiro. Cada medida univariada pode ter distribuição normal, mas o conjunto de dados inteiro ainda assim pode não ter uma distribuição normal multivariada (Looney, 1995). Os testes adicionais para normalidade multivariada devem ser feitos mesmo se cada variável individual tiver distribuição normal univariada.

Um teste para normalidade multivariada baseia-se em medidas multivariadas de assimetria e curtose (*ver* Capítulo 3). Esse teste foi desenvolvido por Mardia (1970) e estendido por Doornik & Hansen (1994). Os cálculos envolvidos no teste de Doornik & Hansen são comparativamente simples e o algoritmo é fornecido por completo no artigo destes autores de 1994.[4]

Testamos os dados de *Darlingtonia* (Tabela 12.1) para a normalidade multivariada usando o teste de Doornik & Hansen. Apesar de todas as medidas

[4] A extensão de Doornik & Hansen (1994) é mais acurada (dá o erro Tipo I esperado nas simulações) e tem um poder estatístico melhor do que o teste de Mardia tanto para tamanhos amostrais pequenos ($50 > N > 7$) quanto grandes ($N > 50$) (Doornik & Hansen, 1994, Mecklin & Mundfrom, 2003). Também é mais fácil de entender e programar do que o procedimento alternativo descrito por Royston (1983).

Para um conjunto de dados multivariado com m observações e n medidas, a estatística do teste de Doornik & Hansen, E_n, é calculada como $E_n = \mathbf{Z}_1^T\mathbf{Z}_1 + \mathbf{Z}_2^T\mathbf{Z}_2$, onde Z_1 é um vetor coluna n-dimensional de assimetria transformada dos dados multivariados e Z_2 é um vetor coluna n-dimensional de curtose transformada dos dados multivariados. As transformações são dadas em Doornik & Hansen (1994). A estatística teste E_n é distribuída assintoticamente como uma variável aleatória χ^2 com $2n$ graus de liberdade. (AME escreveu uma função em S-Plus/R para realizar o teste de Doornik & Hansen. O código pode ser baixado em http://harvardforest.fas.harvard.edu/personnel/web/aellison.)

individuais terem distribuição normal ($P > 0,5$ para todas variáveis, usando o teste de bondade do ajuste de Kolmogorov-Smirnov), os dados multivariados pouco desviaram da normalidade multivariada ($P = 0,006$). Essa falta de ajuste foi consequência da medida da largura da quilha do tubo. Removendo essa variável, os dados remanescentes passaram no teste de normalidade multivariada ($P = 0,07$).

MEDIDAS DE DISTÂNCIA MULTIVARIADA

Vários métodos multivariados quantificam a diferença entre observações individuais, amostras, grupos de tratamentos ou populações. Essas diferenças com frequência são expressas como distâncias entre observações no espaço multivariado. Antes de mergulharmos nas técnicas analíticas, descrevemos como as distâncias são calculadas entre observações individuais ou entre os centroides que representam as médias de grupos inteiros. Usando os dados amostrais da Tabela 12.1, poderíamos perguntar: quão distante (ou diferentes) duas plantas individuais são uma da outra no espaço multivariado criado usando as variáveis morfológicas?

Medindo distâncias entre dois indivíduos

Vamos começar considerando apenas dois indivíduos, planta 1 e planta 2, da Tabela 12.1, e duas das variáveis, altura do jarro (variável 1) e envergadura do seu apêndice cauda de peixe (variável 2). Para essas duas plantas, poderíamos medir a distância morfológica entre elas plotando os pontos em duas dimensões e medindo a menor distância entre eles (Figura 12.2). Essa distância, obtida aplicando o teorema de Pitágoras,[5] é chamada de **distância euclidiana** e é calculada como:

[5] Pitágoras de Samos (ca. 569-475 a.C.) deve ter sido o primeiro matemático "puro". Ele é mais bem lembrado por ser o primeiro a oferecer uma prova formal do que agora é conhecido como o Teorema de Pitágoras: a soma dos quadrados dos catetos de um triângulo reto é igual ao quadrado do comprimento da sua hipotenusa: $a^2 + b^2 = c^2$. Note que a distância euclidiana que estamos usando é o comprimento da hipotenusa de um triângulo reto implícito (*ver* Figura 12.2). Pitágoras também sabia que a Terra era esférica e, embora a colocando no centro do universo, ele errou sua posição por alguns bilhões de anos-luz da sua posição real.

Pitágoras

Pitágoras fundou uma sociedade científico-religiosa em Crotona, atualmente sul da Itália. A panelinha dessa sociedade (que incluía homens e mulheres) era constituída os *mathematikoi*, comunitários que renunciaram as posses pessoais e eram estritamente vegetarianos. Entre as suas crenças fundamentais estava a de que a realidade é natureza matemática, que essa filosofia é a base da purificação espiritual, que a alma pode obter a união com o divino e que alguns símbolos têm significado místico. Todos os membros da ordem eram jurados para lealdade e segredo máximos, por isso dados biográficos sobre Pitágoras são escassos.

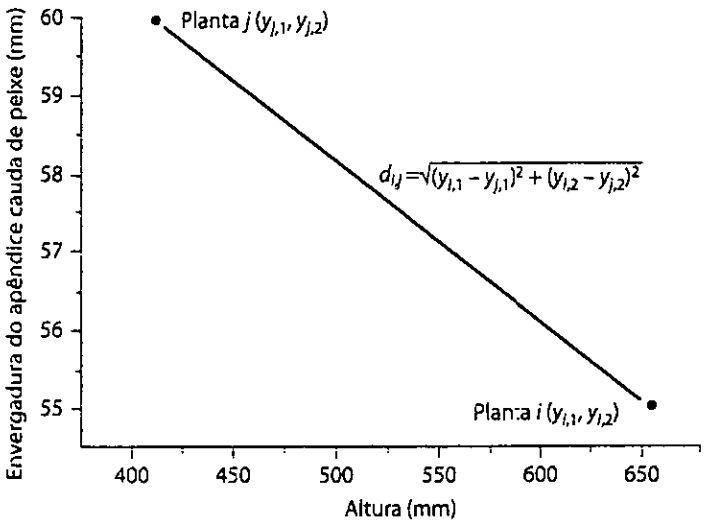


Figura 12.2 Em duas dimensões, a distância euclidiana d_{ij} é aquela em linha reta entre os pontos. Neste exemplo, as duas variáveis morfológicas que medimos são a altura da planta e a envergadura do apêndice da extremidade do lírio-cobra carnívoro, *Darlingtonia californica* (Tabela 12.1). Aqui, as medidas para duas plantas individuais são plotadas no espaço bivariado. A Equação 12.15 é usada para calcular a distância euclidiana entre elas.

$$d_{i,j} = \sqrt{(y_{i,1} - y_{j,1})^2 + (y_{i,2} - y_{j,2})^2} \qquad (12.15)$$

Na Equação 12.15, a planta é indicada pela letra subscrita i ou j, e a variável pelo subscrito 1 ou 2. Para calcular a distância entre as medidas, elevamos ao quadrado a diferença entre as alturas, somamos a isto o quadrado da diferença das envergaduras e tiramos a raiz quadrada dessa soma. O resultado para essas duas plantas é a distância $d_{i,j} = 241{,}05$ mm. Como as variáveis originais foram medidas em milímetros, quando subtraímos uma medida de outra, nosso resultado ainda está em milímetros. Aplicando as operações subsequentes de elevar ao quadrado, somar e tirar a raiz quadrada traz de volta uma medida de distância que ainda está nas unidades de milímetros. Entretanto, as unidades de medidas de distância não permanecerão as mesmas, a menos que todas as variáveis originais forem medidas nas mesmas unidades.

Similarmente, podemos calcular a distância euclidiana entre essas duas plantas se elas forem descritas por três variáveis: altura, envergadura e diâmetro da boca (Figura 12.3). Apenas adicionamos outra diferença elevada ao quadrado – neste caso, existente entre os diâmetros da boca – à Equação 12.15:

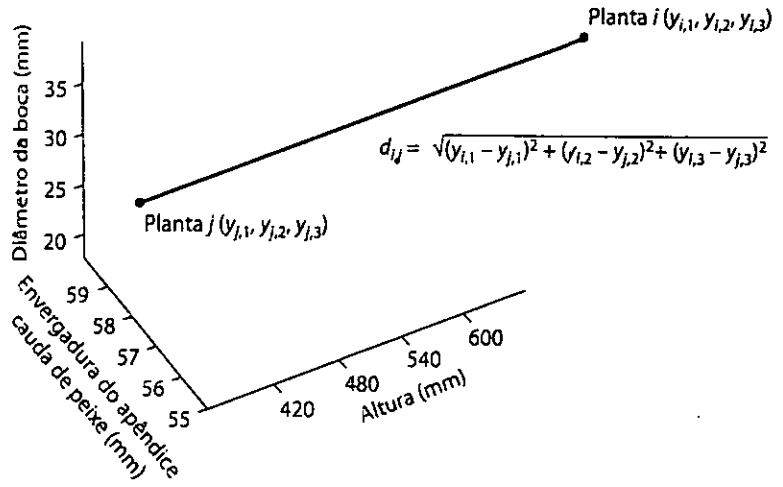


Figura 12.3 Medindo a distância euclidiana em um espaço tridimensional. O diâmetro da boca foi adicionado às duas variáveis morfológicas mostradas na Figura 12.2, e as três variáveis são plotadas no espaço tridimensional. A Equação 12.16 é usada para calcular a distância euclidiana, que é 241,58. Note que essa distância é virtualmente idêntica à euclidiana medida em duas dimensões (241,05, Figura 12.2). A razão é que a terceira variável (diâmetro da boca) tem média e variância muito menores que as duas primeiras variáveis, e então não afeta muito a medida de distância. Por essa razão, as variáveis devem ser padronizadas (usando a Equação 12.17) antes de calcular a distância entre indivíduos.

$$d_{i,j} = \sqrt{(y_{i,1} - y_{j,1})^2 + (y_{i,2} - y_{j,2})^2 + (y_{i,3} - y_{j,3})^2} \qquad (12.16)$$

O resultado da aplicação da Equação 12.16 às duas plantas é uma distância morfológica de 241,58 mm – uma mera mudança de 0,2% da distância obtida usando apenas as duas variáveis altura e envergadura.

Por que essas duas distâncias euclidianas não são muito diferentes? Pois a magnitude da altura da planta (centenas de milímetros) é muito maior que a magnitude tanto da envergadura quanto do diâmetro da boca (dezenas de milímetros), a medida da altura da planta domina os cálculos da distância. Na prática, portanto, é essencial padronizar as variáveis antes de calcular distâncias. Uma padronização conveniente é transformar cada variável subtraindo sua média amostral do valor de cada observação daquela variável, e então dividir essa diferença pelo desvio-padrão amostral:

$$Z = \frac{(Y_i - \overline{Y})}{s} \qquad (12.17)$$

O resultado dessa transformação é chamado **escore-*Z***. Seu papel é controlar as diferenças em cada variável medida e também comparar medidas que não estão nas mesmas unidades. Os valores padronizados (escores-*Z*) da Tabela 12.1 são dados na Tabela 12.3. A distância entre essas duas plantas para as duas variáveis padronizadas, altura e envergadura, é 2,527, e para as três variáveis padronizadas, altura, envergadura e diâmetro da boca, é 2,533. A diferença absoluta entre essas duas distâncias é de modestos 1%. Contudo, essa diferença é cinco vezes maior que a existente entre as distâncias calculadas com os dados não padronizados.

Quando os dados foram transformados com os escores-*Z*, quais são as unidades de distância? Se estivéssemos trabalhando com as variáveis originais, as unidades seriam em milímetros. Mas variáveis padronizadas sempre são adimensionais. A Equação 12.17 fornece unidades de mm × mm^{-1}, que se cancelam. Normalmente, pensamos nos escores-*Z* como parte de unidades de "desvios-padrões" – quantos desvios padrões uma medida está da média. Por exemplo, uma planta com uma medida de altura padronizada de 0,381 é 0,381 desvios-padrões maior do que a planta média.

Embora não possamos desenhar quatro ou mais eixos em uma folha de papel plana, podemos calcular, ainda assim, as distâncias euclidianas entre dois indivíduos se eles são descritos por n variáveis. A fórmula geral para a distância euclidiana com base em um conjunto de n variáveis é:

$$d_{i,j} = \sqrt{\sum_{k=1}^{n} (y_{i,k} - y_{j,k})^2} \qquad (12.18)$$

Conforme todas as 7 variáveis morfológicas padronizadas, a distância euclidiana entre as duas plantas na Tabela 12.3 é 4,346.

TABELA 12.3 Padronização de dados multivariados

Local	Planta	Altura	Boca	Tubo	Quilha	Asa1	Asa2	Envergadura	Massa do capuz	Massa do tubo	Massa da asa
TJH	1	0,381	1,202	–1,062	0,001	0,516	0,156	–1,035	1,568	0,602	0,159
TJH	2	–2,014	–1,388	–0,878	–0,228	–0,802	–1,984	–0,893	–0,881	–1,281	–0,276
A	A	A	A	A	A	A	A	A	A	A	A

Os dados originais são medidas de 7 variáveis morfológicas e 3 variáveis de biomassa de 89 indivíduos de *Darlingtonia californica* coletados em quatro locais (Tabela 12.1). As duas primeiras linhas de dados são ilustradas após a padronização. Os valores padronizados são calculados subtraindo cada observação da média amostral de cada variável e dividindo essa diferença pelo desvio-padrão amostral (Equação 12.17).

Medindo distâncias entre dois grupos

Tipicamente, queremos medir distâncias entre amostras ou grupos de tratamentos experimentais, não apenas entre indivíduos. Podemos aplicar a fórmula da distância euclidiana para medir essa distância entre médias de qualquer número arbitrário de grupos g:

$$d_{i,j} = \sqrt{\sum_{k=1}^{g} (\overline{Y}_{i,k} - \overline{Y}_{j,k})^2} \qquad (12.19)$$

onde $\overline{Y}_{i,k}$ é a média da variável i no grupo k. O Y_i, na Equação 12.19, pode ser univariado ou multivariado. Mais uma vez, para que nenhuma variável domine o cálculo de distância, em geral padronizamos os dados (Equação 12.17) antes de calcular as médias e as distâncias entre os grupos. A partir da Equação 12.19, a distância euclidiana entre as populações de DG e TJH, fundamentada em todas as 7 variáveis morfológicas, é de 1,492 desvios-padrões.

Outras medidas de distância

A distância euclidiana é a medida mais usada. Entretanto, ela nem sempre é a melhor escolha para calcular o afastamento entre objetos multivariados. Por exemplo, uma questão comum em ecologia de comunidades é quão diferente dois locais são com base na ocorrência ou abundância das espécies encontradas nesses dois locais. Um paradoxo curioso e contraintuitivo é que dois locais com nenhuma espécie em comum podem ter uma distância euclidiana menor que outros dois que compartilham pelo menos algumas espécies! Esse paradoxo é ilustrado usando dados hipotéticos (Orlóci, 1978) para três locais $\mathbf{x}_1$, $\mathbf{x}_2$, e $\mathbf{x}_3$, e três espécies

TABELA 12.4 Matriz local × espécies

	Espécies		
Local	y_1	y_2	y_3
x_1	0	1	1
x_2	1	0	0
x_3	0	4	4

Essa matriz ilustra o paradoxo de se usar distâncias euclidianas para medir a similaridade entre locais com base na abundância de espécies. Cada linha representa um local e cada coluna, uma espécie. Os valores são os números de indivíduos de cada espécie y_i no local x_j. O paradoxo é que a distância euclidiana entre dois locais sem nenhuma espécie em comum (como os locais x_1 e x_2) pode ser menor que a euclidiana entre dois lugares que compartilham pelo menos algumas espécies (como x_1 e x_3).

TABELA 12.5 Distâncias euclidianas par-a-par entre os locais

	Local		
Local	x_1	x_2	x_3
x_1	0	1,732	4,243
x_2	1,732	0	5,745
x_3	4,243	5,745	0

Essa matriz baseia-se nas abundâncias das espécies dadas na Tabela 12.4. Cada valor na matriz corresponde à distância euclidiana entre os locais x_i e x_k. Note que a matriz de distância é simétrica (*ver* o Apêndice): por exemplo, a distância entre os locais x_1 e x_2 é a mesma que a existente entre x_2 e x_1. Os elementos da diagonal são todos iguais a 0, porque a distância entre um local e ele mesmo é = 0. Essa matriz de cálculos de distância demonstra que as euclidianas podem dar resultados contraintuitivos se os dados contêm muitos zeros. Os locais x_1 e x_2 não partilham espécies em comum, mas sua distância euclidiana é inferior à mesma existente entre x_1 e x_3, que compartilham o mesmo conjunto de espécies.

y_1, y_2, y_3. A Tabela 12.4 dá a matriz de locais × espécies; a Tabela 12.5, todas as distâncias euclidianas par-a-par entre os locais. Neste exemplo simples, os locais x_1 e x_2 não têm espécies em comum, e a distância euclidiana entre eles é 1,732 espécies (Equação 12.14):

$$d_{x_1, x_2} = \sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = 1{,}732$$

Em contraste, os locais x_1 e x_3 têm todas suas espécies em comum (tanto y_2 quanto y_3 ocorrem nessas duas áreas), mas a distância entre eles é 4,243 espécies:

$$d_{x_1, x_3} = \sqrt{(0-0)^2 + (1-4)^2 + (1-4)^2} = 4{,}243$$

Os ecólogos, cientistas da área ambiental e estatísticos desenvolveram muitas outras medidas para quantificar a distância entre duas amostras ou populações multivariadas. Um subconjunto desses é apresentado na Tabela 12.6, Legendre & Legendre (1998) e Podani & Miklós (2002) discutem muitas outras. Essas medidas dividem-se em duas categorias: **distâncias métricas** e **distâncias semi-métricas**.

As distâncias métricas têm quatro propriedades:

1. A distância mínima é = 0, e se dois objetos (ou amostras) x_i e x_2 são idênticos, então a distância d entre eles também é igual a 0: $x_1 = x_2 \Rightarrow d(x_1,x_2) = 0$.
2. A medida de distância d é sempre positiva se dois objetos (x_1 e x_2) não são idênticos: $x_1 \neq x_2 \Rightarrow d(x_1,x_2) > 0$.
3. A medida de distância d é simétrica: $d(x_1,x_2) = d(x_2,x_1)$.
4. A medida de distância d satisfaz a **desigualdade triangular**: para três objetos x_1, x_2 e x_3, $d(x_1,x_2) + d(x_2,x_3) \geq d(x_1,x_3)$.

TABELA 12.6 Algumas medidas comuns de distância ou dissimilaridade usadas por ecólogos

Nome	Fórmula	Propriedade
Euclidiana	$d_{i,j} = \sqrt{\sum_{k=1}^{n} (y_{i,k} - y_{j,k})^2}$	Métrica
Manhattan (ou City Block)	$d_{i,j} = \sum_{k=1}^{n} \lvert y_{i,k} - y_{j,k} \rvert$	Métrica
Corda	$d_{i,j} = \sqrt{2 \times \left(1 - \frac{\sum_{k=1}^{n} y_{i,k} y_{j,k}}{\sqrt{\sum_{k=1}^{n} y_{i,k}^2 \sum_{k=1}^{n} y_{j,k}^2}}\right)}$	Métrica
Mahalanobis	$d_{\mathbf{y}_i,\mathbf{y}_j} = \mathbf{d}_{i,j}\mathbf{V}^{-1}\mathbf{d}_{i,j}^{\mathrm{T}}$ $\mathbf{V} = \frac{1}{m_i + m_j - 2}[(m_i - 1)\mathbf{C}_i + (m_j - 1)\mathbf{C}_j]$	Métrica
Chi-quadrado	$d_{i,j} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} y_{ij}} \times \sqrt{\sum_{k=1}^{n}\left[\frac{1}{\sum_{k=1}^{n} y_{jk}} \times \left(\frac{y_{ik}}{\sum_{k=1}^{n} y_{ik}} - \frac{y_{jk}}{\sum_{k=1}^{n} y_{jk}}\right)^2\right]}$	Métrica
Bray-Curtis	$d_{i,j} = \frac{\sum_{k=1}^{n} \lvert y_{i,k} - y_{j,k} \rvert}{\sum_{k=1}^{n} (y_{i,k} + y_{j,k})}$	Semimétrica
Jaccard	$d_{i,j} = \frac{a+b}{a+b+c}$	Métrica
Sørensen	$d_{i,j} = \frac{a-b}{a+b+2c}$	Semimétrica

As distâncias euclidiana, de Manhattan (ou City Block), de Corda, e de Bray-Curtis são usadas principalmente para dados numéricos contínuos. As de Jaccard e Sørensen são usadas para medir o afastamento entre duas amostras que são descritas por dados de presença e ausência. Nas de Jaccard e Sørensen, a é o número de objetos (p. ex., espécies) que ocorrem apenas em y_i, b é o número de objetos que ocorrem apenas em y_j e c é o número de objetos que ocorre em ambos, em y_i e em y_j. A distância de Mahalanobis se aplica apenas para *grupos* de amostras $\mathbf{y}_i$ e $\mathbf{y}_j$, em que cada um contêm m_i e m_j amostras, respectivamente. Na equação da distância de Mahalanobis, $\mathbf{d}$ é o vetor de diferenças entre as médias das m amostras em cada grupo, $\mathbf{V}$ é a matriz conjunta de variância e covariância amostral dentro de grupos, calculada como demonstrado, e $\mathbf{C}_i$ é a matriz de variância e covariância amostral para $\mathbf{y}_i$ (Equação 12.3).

As distâncias euclidiana, de Manhattan, de Chord, de Mahalanobis, chi quadrado e de Jaccard[6] são exemplos de distâncias métricas.

As distâncias semimétricas satisfazem apenas as três primeiras dessas propriedades, mas podem violar a desigualdade triangular. As medidas de Bray-Curtis e de Sørensen são semimétricas. Um terceiro tipo de medida de distância, não usa-

[6] Se você já tiver usado o índice similaridade de Jaccard, pode perguntar por que nos referimos a ele na Tabela 12.6 como uma medida de distância ou dissimilaridade. O coeficiente de Jaccard (1901) foi desenvolvido para descrever quão similares duas comunidades são em termos de espécies compartilhadas $s_{i,j}$

$$s_{i,j} = \frac{c}{a+b+c}$$

onde a é o número de espécies que ocorrem apenas na comunidade i, b é o número de espécies que ocorre apenas na comunidade j, e c é o número de espécies que elas têm em comum. Como as medidas de similaridade assumem seu valor máximo quando os objetos são mais similares, e as medidas de dissimilaridade (ou distância) assumem seu valor máximo quando os objetos são mais diferentes, qualquer medida de similaridade pode ser transformada em uma medida de dissimilaridade ou distância. Se uma medida de similaridade s varia de 0 a 1 (como o coeficiente de Jaccard), ela pode ser transformada em uma medida de distância d usando umas das três equações a seguir:

$$d = 1 - s, d = \sqrt{1-s}, \text{ou } d = \sqrt{1-s^2}$$

Deste modo, na Tabela 12.6, a distância de Jaccard é igual a 1 – o coeficiente de similaridade de Jaccard. A transformação reversa (p. ex., $s = 1 - d$) pode ser usada para transformar medidas de distância em medidas de similaridade. Se a medida de distância não é delimitada (i. e., varia de 0 a ∞), ela deve ser normalizada para variar entre 0 e 1:

$$d_{norm} = \frac{d}{d_{\max}} \text{ ou } d_{norm} = \frac{d - d_{\min}}{d_{\max} - d_{\min}}$$

Embora essas propriedades algébricas dos índices de similaridade sejam simples, suas propriedades estatísticas não são. Índices de similaridade, como usados na biogeografia e ecologia de comunidades, são muito sensíveis à variação no tamanho amostral (Wolda, 1981; Jackson et al., 1989). Em particular, pequenas comunidades podem ter coeficientes de similaridade relativamente altos mesmo na ausência de qualquer força biótica incomum, porque suas faunas são dominadas por um punhado de espécies comuns e com ampla distribuição. Espécies raras são encontradas principalmente em faunas maiores, e elas tenderão a reduzir o índice de similaridade entre duas comunidades mesmo que as espécies raras também ocorram de forma aleatória. Os índices de similaridade e outros índices bióticos deveriam ser comparados a um modelo nulo apropriado que controle a variação no tamanho amostral ou o número de espécies presente (Gotelli & Graves, 1996). Tanto simulações de Monte Carlo quanto cálculos de probabilidade podem ser usados para determinar o valor esperado de um índice de biodiversidade ou similaridade em pequenos tamanhos amostrais (Colwell & Coddington, 1994; Gotelli & Colwell, 2001). A similaridade medida na composição de espécies de duas assembleias depende não apenas do número de espécies que elas compartilham, mas do potencial de dispersão das espécies (a maioria dos modelos simples assume que as ocorrências das espécies são equiprováveis), e da composição e do tamanho do *pool* de espécies (Connor & Simberloff, 1978; Rice & Belland, 1982). Alternativamente, Legendre & Gallagher (2001) sugerem transformações que reduzem a influência das espécies raras nesses tipos de comparações.

da por ecólogos, é a **não métrica**, que viola a segunda propriedade das distâncias métricas e pode assumir valores negativos.

ORDENAÇÃO

As técnicas de **ordenação** são usadas para ou ordenar dados multivariados. A ordenação cria novas variáveis (chamadas **eixos principais**), através das quais as amostras recebem escores ou são colocadas em ordem. Esse ordenamento pode representar uma simplificação útil de padrões em conjuntos de dados multivariados complexos. Usada dessa maneira, a ordenação é uma técnica de redução de dados: começando com um conjunto de n variáveis, gera um número menor de variáveis que, ainda assim, ilustram os padrões importantes nos dados. A ordenação também pode ser usada para discriminar ou separar amostras ao longo de eixos.

Os ecólogos e cientistas da área ambiental muitas vezes usam cinco tipos diferentes de ordenação – análise de componentes principais, fatorial, de correspondência, de coordenadas principais e escalonamento multidimensional não métrico – que discutiremos em separado. Avaliamos profundamente os detalhes da análise de componentes principais, porque os conceitos e métodos são similares àqueles usados em outras técnicas de ordenação. Nossa discussão sobre os outros quatro métodos é um pouco breve. Legendre & Legendre (1998) é um bom guia ecológico para os detalhes da análise multivariada.

Análise de componentes principais

A **análise de componentes principais** (*Principal Component Analysis* – PCA) é a maneira mais simples para ordenar dados. A ideia da PCA geralmente é creditada a Karl Pearson (1901),[7] o renomado Pearson do chi-quadrado e da correlação, porém Harold Hotelling[8] (do teste T^2 de Hotelling) desenvolveu os métodos computacionais em 1933. O uso básico da PCA reduz a dimensionalidade de dados multivariados. Em outras palavras, usamos a PCA para criar algumas poucas va-

[7] O objetivo de Pearson no seu artigo de 1901 era determinar a atribuição racial de indivíduos com base em múltiplas medidas biométricas. *Ver*, na Nota de Rodapé 3, Capítulo 11, um esboço biográfico de Karl Pearson.

Harold Hotelling

[8] Harold Hotelling (1895-1973) contribuiu significativamente para a estatística e para a economia, dois campos dos quais os ecólogos surrupiaram muitas ideias. Seu artigo de 1931 estendendo o teste-t para casos multivariados também introduziu os intervalos de confiança para os estatísticos, e o seu artigo de 1933 desenvolveu a matemática dos componentes principais. Como economista, ele era mais conhecido por defender abordagens com base em custos e benefícios marginais – a tradição neoclássica na economia fundamentada no *Manual da Economia Política* de Vilfredo Pareto. Essa tradição também forma a base para vários modelos de otimização do balanço (*trade-offs*) ecológico e evolutivo.

riáveis-chave (onde cada uma é composta de muitas de nossas variáveis originais) que caracterizem o máximo possível a variação em um conjunto de dados multivariado. O atributo mais importante da PCA é que as novas variáveis não são correlacionadas umas com as outras. Essas variáveis não correlacionadas podem ser usadas na regressão múltipla (Capítulo 9) ou na ANOVA (Capítulo 10) sem temer a multicolinearidade. Se as variáveis originais têm distribuição normal, ou tiverem sido transformadas (Capítulo 8) ou padronizadas (Equação 12.17) antes da análise, as novas variáveis resultantes da PCA também terão distribuição normal, satisfazendo um dos requerimentos básicos dos testes paramétricos que são usados para testar hipóteses.

PCA: O CONCEITO A Figura 12.4 ilustra o conceito básico de uma análise de componentes principais. Imagine que você contou o número de indivíduos de duas espécies de tico-tico-do-campo em cada uma de dez parcelas de pradaria. Esses dados podem ser plotados em um gráfico no qual a abundância da Espécie A está no eixo-x e a da Espécie B, no eixo-y. Cada ponto no gráfico representa os dados de uma parcela diferente. Agora, criamos uma nova variável, ou **primeiro eixo principal**,[9] que passa através do centro da nuvem de dados. Em seguida, calculamos um valor para cada uma das dez parcelas ao longo desse novo eixo, e então o representamos. Para cada ponto, esse cálculo usa os números da Espécie A e da Espécie B, gerando um único novo valor, que é o **escore do componente principal** no primeiro eixo principal. Embora os dados multivariados originais tenham duas observações para cada parcela, o escore do componente principal reduz essas duas observações para um único número. Efetivamente, reduzimos a dimensionalidade dos dados de duas dimensões (abundância da Espécie A e da Espécie B) para uma dimensão (primeiro eixo principal).

Em geral, se você mediu de $j = 1$ a n variáveis Y_j para cada réplica, pode gerar n novas variáveis Z_j, que são todas não correlacionadas umas com as outras (todos os elementos fora da diagonal em $\mathbf{P} = 0$). Fazemos isso por duas razões. Primeiro, como os Z_j não são correlacionados, cada um deles pode ser imaginado como medindo uma "dimensão" diferente e independente dos dados multivariados. Note que essas novas variáveis não são as mesmas que os escores-Z descritos acima (Equação 12.17).

Segundo, as novas variáveis podem ser ordenadas com base na quantidade de variação que elas explicam nos dados originais. Então, Z_1 tem a maior quantidade de variação nos dados (e é chamado **eixo maior**), Z_2 tem a próxima maior

[9] O termo "eixo principal" é derivado da óptica e é usado para se referir à linha (também conhecida como eixo óptico) que passa através do centro da curvatura de uma lente, de modo que um raio de luz passando ao longo dessa linha não é nem refletido nem refratado. Também é usado na física para se referir a um eixo de simetria, em que um objeto irá rotacionar em uma velocidade angular constante sobre o seu eixo principal, sem a aplicação de torque adicional.

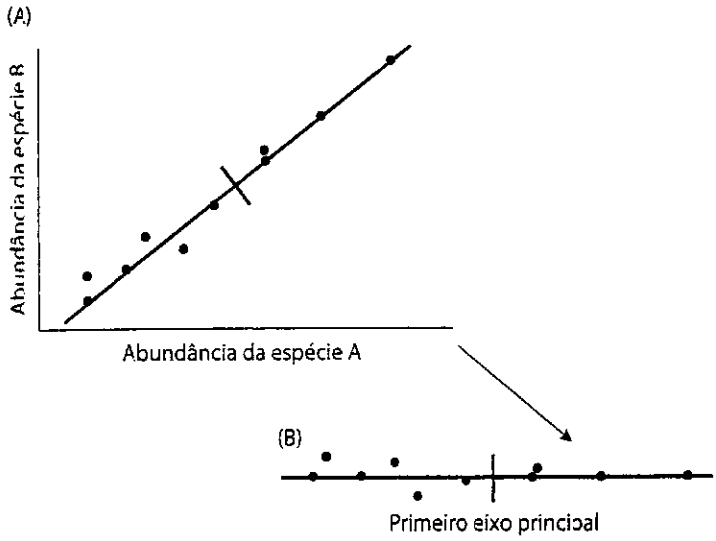


Figura 12.4 Construção de um eixo de componente principal. (A) Para um conjunto hipotético de 10 parcelas de pradaria, o número de indivíduos de duas espécies de tico-tico-do-campo é medido em cada parcela. Cada uma pode ser representada como um ponto em um gráfico de dispersão bivariado com a abundância de cada espécie representada nos dois eixos. O primeiro eixo principal (linha verde) passa através do maior eixo de variação dos dados. O segundo componente principal (pequena linha preta) é ortogonal (perpendicular) ao primeiro e explica a pequena quantidade de variação residual não incorporada anteriormente no primeiro eixo principal. (B) O primeiro eixo principal é então rotacionado, se tornando o novo eixo ao longo do qual os escores das parcelas são colocados em ordem (ordenados). A linha preta vertical indica a interseção do segundo eixo principal. Pontos à direita da linha vertical têm escores positivos no primeiro eixo principal, e pontos à esquerda da linha vertical têm escores negativos. Apesar de os dados originais multivariados terem duas medidas para cada parcela, a maioria da variação nos dados é capturada por um único escore no primeiro componente principal, permitindo que as parcelas sejam ordenadas ao longo desse eixo.

quantidade de variação nos dados, e assim por diante ($\text{var}(Z_1) \geq \text{var}(Z_2) \geq \ldots \geq \text{var}(Z_n)$). Em um típico conjunto de dados ecológico, a maioria da variação nos dados é capturada pelos primeiros Z_j, e podemos descartar os subsequentes, que explicam a variação residual. Ordenados dessa maneira, os Z_j são chamados **componentes principais**. Se a PCA é informativa, reduz um grande número de variáveis originais correlacionadas para um pequeno número de variáveis novas e não correlacionadas.

A análise de componentes principais é bem-sucedida na medida em que existem fortes intercorrelações nos dados originais. Se no começo todas as variáveis não são correlacionadas, não ganhamos nada com a PCA, pois não podemos capturar a variação em um punhado de variáveis novas e transformadas; podemos, sim, basear a análise nas próprias variáveis não transformadas. Além disso, a PCA não será bem-sucedida se empregada em variáveis não pa-

dronizadas. A padronização é necessária para a mesma escala relativa, usando transformações como a da Equação 12.17, de modo que os eixos da PCA não sejam dominados por uma ou duas variáveis que tenham grandes unidades de medida.

UM EXEMPLO DE UMA PCA Então como é que funciona? A título de exemplo, vamos examinar alguns dos resultados de uma PCA aplicada aos dados de *Darlingtonia* na Tabela 12.1. Para esta análise, usamos todas as dez variáveis – altura, abertura da boca, diâmetros do tubo e da quilha, comprimento e envergadura das asas do apêndice cauda de peixe, e massas do capuz, do tubo e do apêndice – que chamaremos de Y_1 a Y_{10}. O primeiro componente principal, Z_1, é uma nova variável (não entre em pânico, os detalhes virão a seguir) cujo valor para cada observação é:

$$Z_1 = 0{,}31Y_1 + 0{,}40Y_2 - 0{,}002Y_3 - 0{,}18Y_4 + 0{,}39Y_5 + 0{,}37Y_6 + 0{,}26Y_7 + 0{,}40Y_8 + 0{,}38Y_9 + 0{,}23Y_{10} \quad \textbf{(12.20)}$$

Geramos o escore do primeiro componente principal (Z_1) para cada planta individual multiplicando cada variável medida vezes o coeficiente correspondente (a **carga**) e somando os resultados. Deste modo, multiplicamos 0,31 vezes a altura da planta, mais 0,40 vezes o diâmetro da abertura da boca, e assim por diante. Para a primeira réplica da Tabela 12.1, $Z_1 = 1{,}43$. Como os coeficientes para alguns dos Y_j podem ser negativos, o escore do primeiro componente para uma réplica particular pode ser positivo ou negativo. A variância de Z_1 é = 4,49 e explica 46% da variação total nos dados. Os segundo e terceiro componentes principais têm variâncias de 1,63 e 1,46, e a de todos os outros é < 0,7.

De onde veio a Equação 12.20? Note primeiro que Z_1 é uma **combinação linear** das dez variáveis Y. Como descrevemos acima, multiplicamos cada variável Y_j por um coeficiente a_{ij}, também conhecido como carga (*loading* em inglês), e somamos todos esses produtos para obter o Z_1. Cada componente principal adicional é outra combinação linear de todos os Y_j:

$$Z_j = a_{i1}Y_1 + a_{i2}Y_2 + \ldots + a_{in}Y_n \quad \textbf{(12.21)}$$

Os a_{ij} são os coeficientes para o fator i que são multiplicados pelo valor da variável j medido. Cada componente principal explica tanta variação nos dados quanto possível, sujeito à condição de que todos os Z_j não sejam correlacionados. Essa condição garante que os Z_j sejam independentes e ortogonais. Portanto, quando plotamos os Z_j, estamos vendo relações (se houver alguma) entre variáveis independentes.

CALCULANDO COMPONENTES PRINCIPAIS Tanto o cálculo dos coeficientes a_{ij}, quanto o de suas variâncias associadas é relativamente simples. Começamos padroni-

zando nossos dados, aplicamos a Equação 12.17 e então calculamos a matriz C de variância e covariância das amostras (Equação 12.3) dos dados padronizados. Note que C calculada usando os dados padronizados é a mesma da matriz de correlação P (Equação 12.5) usando dados brutos.

Calculamos, então, os **autovalores** $\lambda_1 \ldots \lambda_n$ da matriz de variância e covariância amostral e seus **autovetores,** $\mathbf{a}_j$, associados (*ver* o Apêndice para os métodos). O j-ésimo autovalor é a variância de Z_j, e as cargas a_{ij} são os elementos dos autovetores. A soma de todos os autovalores é a variância total explicada:

$$\text{var}_{total} = \sum_{j=1}^{n} \lambda_j$$

e a proporção da variância explicada por cada componente Z_i é:

$$\text{var}_j = \frac{\lambda_j}{\sum_{j=1}^{n} \lambda_j}$$

Se multiplicarmos var_j por 100, obtemos a porcentagem de variação explicada. Como existem tantos componentes principais quantas forem as variáveis originais, a quantidade total de variação explicada por todos os componentes principais é = 100%. Os resultados desses cálculos para os dados da *Darlingtonia* na Tabela 12.1 são resumidos nas Tabelas 12.7 e 12.8.

QUANTOS COMPONENTES USAR? Como a PCA é um método para simplificar dados multivariados, estamos interessados em reter apenas aqueles componentes que explicam a maior parte da variação nos dados. Não existe nenhum ponto de corte absoluto a partir do qual descartamos os componentes principais, mas uma ferramenta gráfica útil para examinar a contribuição de cada componente principal para toda a PCA é um ***scree plot*** (Figura 12.5). Esse gráfico ilustra a porcentagem de variação (o autovalor) em ordem decrescente explicada por cada componente. O *scree plot* parece o perfil de uma montanha abaixo com bastante pedregulho (conhecido como inclinação de pedreira). Os dados usados para construir o *scree plot* estão tabulados na Tabela 12.7. Em um *scree plot*, procuramos por uma curva abrupta ou uma mudança na inclinação dos autovalores ordenados. Retemos aqueles componentes que contribuem para o lado da montanha e ignoramos o cascalho no fundo. Neste exemplo, os Componentes 1-3 parecem úteis, explicando um total de 77% da variação, enquanto os Componentes 4-10 parecem apenas cascalho, e nenhum deles explica mais que 7% da variação restante. Jackson (1993) discute uma ampla variedade de métodos heurísticos (como o *scree plot*) e estatísticos para escolher quantos componentes principais usar. Seus resultados sugerem que os *scree plots* tendem a superestimar em 1, o número de componentes principais estatisticamente significantes.

TABELA 12.7 Autovalores da análise de componentes principais

Componente principal	Autovalor λ_i	Proporção de variância explicada	Proporção cumulativa de variância explicada
Z_1	4,51	0,458	0,458
Z_2	1,65	0,167	0,625
Z_3	1,45	0,148	0,773
Z_4	0,73	0,074	0,847
Z_5	0,48	0,049	0,896
Z_6	0,35	0,036	0,932
Z_7	0,25	0,024	0,958
Z_8	0,22	0,023	0,981
Z_9	0,14	0,014	0,995
Z_{10}	0,05	0,005	1,000

Esses autovalores foram obtidos da análise das medidas padronizadas de 7 variáveis morfológicas e 3 variáveis de biomassa de 89 indivíduos do lírio-cobra *Darlingtonia californica* coletadas em 4 locais (Tabela 12.3). Em uma análise de componentes principais, o autovalor mede a proporção de variância dos dados originais explicada por cada componente principal. A proporção de variância explicada e a proporção cumulativa explicada são calculadas a partir da soma dos autovalores ($\sum\lambda_j = 9{,}83$). A proporção de variância explicada é usada para selecionar um pequeno número de componentes principais que captura a maioria da variação nos dados. Neste conjunto, os três primeiros componentes principais explicam 77% da variância nas 10 variáveis originais.

TABELA 12.8 Autovetores de uma análise de componentes principais

$$\mathbf{a}_1 = \begin{bmatrix} 0{,}31 \\ 0{,}40 \\ -0{,}002 \\ -0{,}18 \\ 0{,}39 \\ 0{,}37 \\ 0{,}26 \\ 0{,}40 \\ 0{,}38 \\ 0{,}23 \end{bmatrix} \quad \mathbf{a}_2 = \begin{bmatrix} -0{,}42 \\ -0{,}025 \\ -0{,}10 \\ -0{,}07 \\ 0{,}28 \\ 0{,}37 \\ 0{,}43 \\ -0{,}18 \\ -0{,}41 \\ 0{,}35 \end{bmatrix} \quad \mathbf{a}_3 = \begin{bmatrix} 0{,}17 \\ -0{,}11 \\ -0{,}74 \\ -0{,}58 \\ -0{,}004 \\ 0{,}09 \\ 0{,}19 \\ -0{,}08 \\ 0{,}04 \\ 0{,}11 \end{bmatrix}$$

Os primeiros três autovetores da análise de componentes principais das medidas da Tabela 12.1. Cada elemento no autovetor é o coeficiente (ou carga) que é multiplicado pelo valor da variável padronizada correspondente (Tabela 12.3). Os produtos das cargas e das medidas padronizadas são somados para dar o escore do componente principal. Desse modo, usando o primeiro autovetor, 0,31 é multiplicado pela primeira medida (altura da planta) e adicionado a 0,40, multiplicado pela segunda medida (diâmetro da boca da planta), e assim por diante. Na notação de matrizes, dizemos que os escores dos componentes principais Z_j para cada observação $\mathbf{y}$, que consiste de dez medidas y_1 até y_{10}, são obtidos multiplicando $\mathbf{a}_j$ por $\mathbf{y}$ (*ver* Equações 12.20 a 12.23).

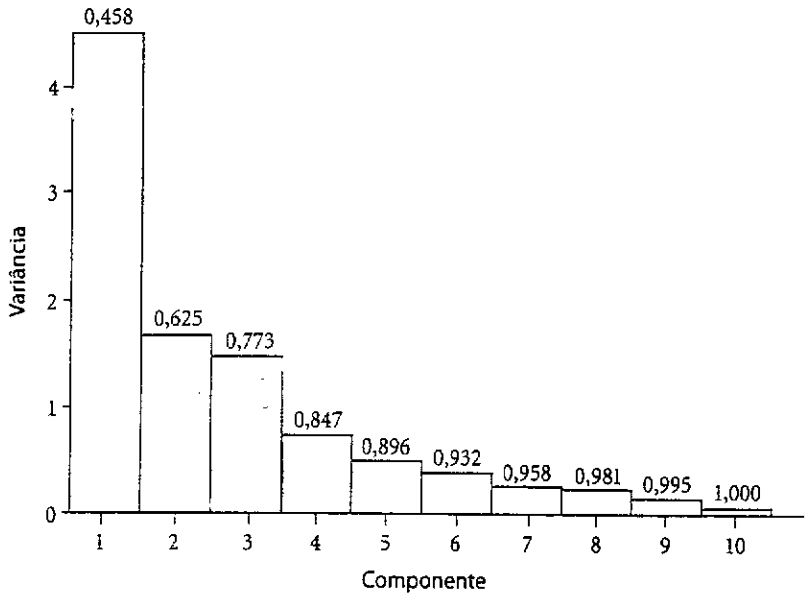


Figura 12.5 Um *scree plot* para a análise de componentes principais dos dados padronizados da Tabela 12.3. Em um *scree plot*, a porcentagem de variância (o autovalor) explicada por cada componente (Tabela 12.7) é apresentada em ordem decrescente. O *scree plot* parece o perfil de uma montanha abaixo, de onde muito pedregulho desmoronou (inclinação de pedreira). Nesse *scree plot*, o Componente 1 claramente domina o lado da montanha, embora os Componentes 2 e 3 também sejam informativos. O restante é cascalho.

O QUE OS COMPONENTES SIGNIFICAM? Agora que selecionamos um pequeno número de componentes, queremos examiná-los com mais detalhes. O Componente 1 já foi descrito na Equação 12.20:

$$Z_1 = 0{,}31Y_1 + 0{,}40Y_2 - 0{,}002Y_3 - 0{,}18Y_4 + 0{,}39Y_5 + 0{,}37Y_6 + 0{,}26Y_7 + 0{,}40Y_8 + 0{,}38Y_9 + 0{,}23Y_{10}$$

Esse componente é a matriz produto do primeiro autovetor $\boldsymbol{a}_1$ (Tabela 12.8) e a matriz de observações multivariadas **Y** que foi padronizada usando a Equação 12.17. A maioria das cargas (ou coeficientes) é positiva, e todas possuem aproximadamente a mesma magnitude, exceto as duas cargas das variáveis relacionadas ao tubo da folha (diâmetro do tubo Y_3 e diâmetro da quilha Y_4). Assim, este primeiro componente Z_1 parece ser uma boa medida do "tamanho" do jarro – plantas com jarros altos, bocas largas e grandes apêndices irão ter maiores valores de Z_1.

Similarmente, o Componente 2 é o produto entre segundo autovetor $\mathbf{a}_2$ (Tabela 12.8) e a matriz de observações padronizadas **Y**:

$$Z_2 = -0{,}42Y_1 - 0{,}25Y_2 - 0{,}10Y_3 - 0{,}07Y_4 + 0{,}28Y_5 + 0{,}37Y_6 + 0{,}43Y_7 - 0{,}18Y_8 - 0{,}41Y_9 + 0{,}35Y_{10} \quad \textbf{(12.22)}$$

Para esse componente, as cargas das seis variáveis relacionadas à altura e ao diâmetro do jarro são negativas, enquanto as cargas das quatro variáveis relacionadas ao apêndice cauda de peixe são positivas. Como todas as variáveis foram padronizadas, os jarros relativamente pequenos e finos terão valores negativos de altura e diâmetro, enquanto os relativamente altos e entroncados terão valores positivos de altura e diâmetro. O valor de Z_2 por consequência reflete *trade-offs* na forma. As plantas pequenas e com grandes apêndices terão grandes valores de Z_2, enquanto as altas com pequenos apêndices terão valores pequenos de Z_2. Podemos pensar em Z_2 como uma variável que descreve a "forma" do jarro. Geramos os escores do segundo componente principal para cada planta individual substituindo de novo seus valores observados y_1 a y_{10} na Equação 12.22.

Por fim, o Componente 3 é o produto entre o terceiro autovetor $\mathbf{a}_3$ (Tabela 12.8) e a matriz de observações padronizadas **Y**:

$$Z_3 = 0{,}17Y_1 - 0{,}11Y_2 + 0{,}74Y_3 + 0{,}58Y_4 - 0{,}004Y_5 + 0{,}09Y_6 + 0{,}19Y_7 - 0{,}08Y_8 - 0{,}04Y_9 + 0{,}11Y_{10} \quad \textbf{(12.23)}$$

Esse componente é significativamente dominado por grandes coeficientes positivos para os diâmetros do tubo (Y_3) e da quilha (Y_4) – medidas de estruturas que sustentam o jarro. O Z_3 será grande para plantas "gordas" e pequeno para as "magras". Como os insetos são capturados dentro do tubo, as plantas com grandes escores Z_3 podem ser capazes de pegar presas maiores do que as com escores menores.

TODAS AS CARGAS SÃO SIGNIFICATIVAS? Ao criar os escores dos componentes principais, usamos as cargas de todas as variáveis originais. Mas algumas são maiores e outras são próximas de zero. Devemos usar todas ou apenas aquelas que estão acima de algum valor de corte? Existe pouco consenso neste ponto (Peres-Neto et al., 2003) e se a PCA é usada apenas para análise exploratória de dados, provavelmente não importa muito se pequenas cargas são retidas ou não. Mas se você usa os escores dos componentes principais para testar hipóteses, deve declarar explicitamente quais cargas foram usadas, e como decidiu por incluí-las ou excluí-las. Se você está testando interações entre os eixos principais, é importante reter todas.

USANDO OS COMPONENTES PARA TESTAR HIPÓTESES Por fim, podemos querer examinar diferenças entre os quatro locais em seus escores dos componentes principais. Podemos ilustrar essas diferenças plotando os escores dos componentes principais de cada planta em dois eixos, e codificar os pontos (= réplicas) para cada grupo usando diferentes cores ou símbolos (Figura 12.6). Neste gráfico, ilustramos os escores dos dois primeiros componentes, Z_1 e Z_2. Poderíamos construir gráficos similares para os outros componentes principais significati-

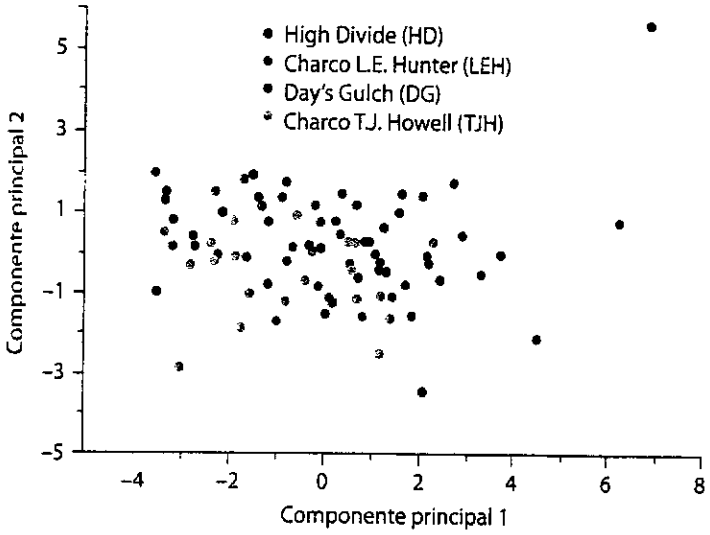


Figura 12.6 Gráfico ilustrando os escores dos dois primeiros componentes principais de uma PCA dos dados na Tabela 12.3. Cada ponto representa uma planta individual e as cores representam os diferentes locais nas montanhas Siskiyou, do sul de Oregon. Embora exista uma sobreposição considerável entre os grupos, também há alguma separação nítida: as amostras de TJH (verde-claro) têm escores mais baixos no componente principal 2, e as amostras de HD (pretas), no primeiro componente principal. Testes univariados ou multivariados podem ser usados para comparar os escores dos componentes principais entre as populações (*ver* Figura 12.7).

vos (p. ex., Z_1 *versus* Z_3; Z_2 *versus* Z_3). A Figura 12.6 foi gerada usando todas as cargas das Equações 12.20 e 12.22.

Uma característica importante da PCA é que as posições dos escores dos componentes principais para cada réplica (como os pontos na Figura 12.6) têm a mesma distância euclidiana entre si que os dados originais no espaço multivariado. Essa propriedade se mantém apenas quando as distâncias euclidianas são calculadas usando *todos* os componentes principais (Z_j). Se você calcular as distâncias euclidianas usando apenas alguns dos primeiros componentes principais, elas não serão iguais às distâncias entre os pontos dos dados originais. Se os dados foram originalmente coletados com informação espacial (como as coordenadas *x*,*y*, ou de latitude e longitude, onde os dados foram coletados), os escores do componentes principais podem ser plotados contra suas coordenadas espaciais. Tais gráficos podem dar informações sobre as relações espaciais multivariadas.

Podemos tratar os escores dos componentes principais como uma variável univariada simples e usar ANOVA (Capítulo 10) para testar as diferenças entre grupos de tratamento ou populações de amostras. Por exemplo, uma ANOVA dos escores do primeiro componente principal dos dados de *Darlingtonia* dá um resultado de que existem diferenças significativas entre os locais ($F_{3,79} = 9{,}26$,

$P = 2 \times 10^{-5}$), e de que todas as diferenças entre pares de locais, exceto para a comparação entre DG e LEH, também são significativamente diferentes (*ver* Capítulo 10 para detalhes dos cálculos de comparações *a posteriori*). Concluímos que o tamanho da planta varia sistematicamente entre locais (Figura 12.7).

Obtemos resultados similares para a forma da planta ($F_{3,79} = 5{,}36$, $P = 0{,}002$), mas apenas três das comparações par-a-par – DG *versus* TJH, HD *versus* TJH e LEH *versus* TJH – são significativamente diferentes. Esse resultado sugere que plantas do TJH têm uma forma diferente das de outros locais. A concentração de pontos verde-claro (os escores dos componentes principais para plantas de TJH) na Figura 12.6 revela que essas plantas têm escores mais baixos para o segundo componente principal – ou seja, são relativamente altas com apêndices cauda de peixe pequenos.

Análise fatorial

A **análise fatorial**[10] e a PCA têm um objetivo similar: a redução de muitas variáveis para poucas variáveis. Enquanto a PCA cria novas variáveis como combinações lineares das variáveis originais (Equação 12.21), a análise fatorial considera cada uma das variáveis originais como uma combinação linear de alguns "fatores" subjacentes. Você pode pensar nisso como uma PCA inversa:

$$Y_j = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{in}F_n + e_j \qquad (12.24)$$

Para usar esta equação, as variáveis Y_j devem ser padronizadas (média = 0, variância = 1) usando a Equação 12.17. Cada a_{ij} é uma **carga do fator**, os F são chamados de **fatores comuns** (e cada um tem média = 0 e variância = 1), e os e_j são fatores específicos da j-ésima variável, que, além disso, não são correlacionados com nenhum F_j, e cada um com média = 0.

Charles Spearman

[10] A análise fatorial foi desenvolvida por Charles Spearman (1863-1945). No seu artigo de 1904, Spearman usou a análise fatorial para medir a inteligência geral a partir da pontuação em testes múltiplos. O modelo fatorial (Equação 12.24) foi usado para a partição dos resultados de uma bateria de "testes de inteligência" (os Y_j) em fatores (os F_j) que mediam a inteligência geral a partir de fatores específicos a cada teste (os e_j). Esta descoberta duvidosa forneceu os fundamentos teóricos para os testes gerais de inteligência de Binet (testes de quociente de inteligência, ou QI). As teorias de Spearman foram usadas para desenvolver os exames britânicos *Plus-Elevens*, que foram administrados em estudantes aos 11 anos de idade no intuito de determinar se eles estariam aptos a frequentar uma universidade ou uma escola técnica ou vocacional. Os testes padronizados em todos os níveis (p. ex., MCAT, SAT e GRE, nos Estados Unidos) são o legado dos trabalhos de Spearman e Binet. O livro *The mismeasure of man** (1981), de Stephen Jay Gould, descreve a história repugnante dos testes de QI e o uso de medidas biológicas a serviço da opressão social.

* N. de T. Título em português: *A falsa medida do homem*.

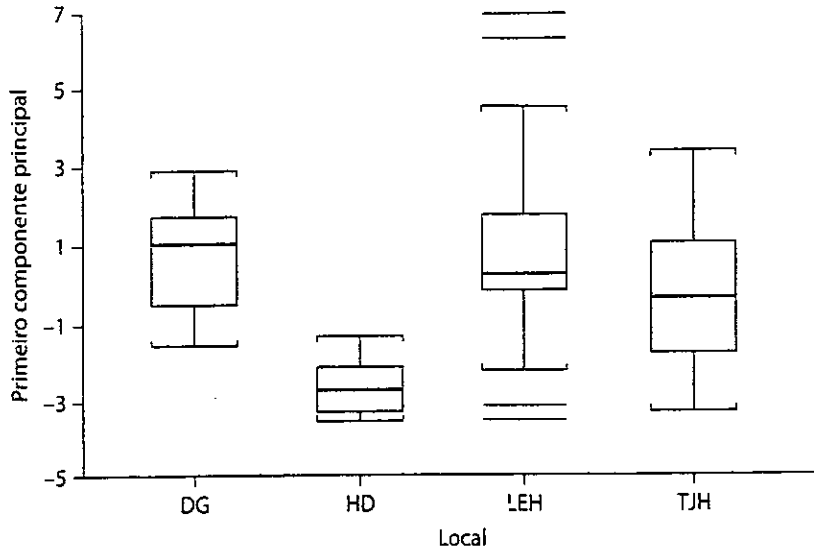


Figura 12.7 *Box plots* dos escores do primeiro componente principal de uma PCA dos dados da Tabela 12.3. Cada *box plot* representa um dos quatro locais especificados na Figura 12.6. A linha horizontal indica a mediana da amostra, e a caixa engloba 50% dos dados, do 25º ao 75º percentil. Os "bigodes" (barras) englobam 90% dos dados, do 10º ao 90º percentil. Quatro pontos de dados extremos (além de 5º e 95º percentis) estão apresentados como barras horizontais para o local LEH. Os escores do Componente Principal 1 variaram de maneira significativa entre os locais ($F_{3,79} = 9{,}26$, $P = 2 \times 10^{-5}$), e foram mais baixos (e menos variáveis) em HD (*ver* Figura 12.6).

Sendo uma "PCA inversa", a análise fatorial geralmente começa com uma PCA e usa os componentes significativos como os fatores iniciais. A Equação 12.21 nos deu a fórmula para gerar componentes principais:

$$Z_j = a_{i1}Y_1 + a_{i2}Y_2 + \ldots + a_{in}Y_n$$

Como a transformação de Y para Z é ortogonal, existe um conjunto de coeficientes a^*_{ij} tal que:

$$Y_j = a^*_{i1}Z_1 + a^*_{i2}Z_2 + \ldots + a^*_{in}Z_n \qquad (12.25)$$

Para uma análise fatorial, mantemos apenas os primeiros m componentes, como aqueles determinados como importantes em uma PCA usando um *scree plot*:

$$Y_j = a^*_{i1}Z_1 + a^*_{i2}Z_2 + \ldots + a^*_{im}Z_m + e_j \qquad (12.26)$$

Para transformar os Z_j em fatores (que são padronizados com variância = 1), dividimos cada um deles pelo seu desvio-padrão, $\sqrt{\lambda_i}$, a raiz quadrada do seu autovalor correspondente. Isso nos dá um **modelo fatorial**:

$$Y_j = b_{i1}F_1 + b_{i2}F_2 + \ldots + b_{im}F_m + e_j \qquad (12.27)$$

onde $F_j = Z_j / \sqrt{\lambda_j}$ e $b_{ij} = a^*_{ij}\sqrt{\lambda_j}$.

ROTACIONANDO OS FATORES Diferente dos Z_j gerados pela PCA, os fatores F_j na Equação 12.27 não são únicos, e mais de um conjunto de coeficientes irá resolver a Equação 12.27. Podemos gerar novos fatores F^*_j, que são combinações lineares dos fatores originais:

$$F^*_j = d_{i1}F_1 + d_{i2}F_2 + \ldots + d_{im}F_m \qquad (12.28)$$

Esses novos fatores também explicam muito da variância dos dados, como os fatores originais. Depois de gerar os fatores usando as Equações 12.25 a 12.27, identificamos os valores para os coeficientes d_{ij} na Equação 12.28, a qual apresenta fatores que são mais fáceis de interpretar. Esse processo de identificação é chamado **rotação** de fatores.

Existem dois tipos de rotação de fatores: a **rotação ortogonal**, que resulta em novos fatores que não são correlacionados uns com os outros, e a **rotação oblíqua**, que resulta em novos fatores que podem ser correlacionados uns com os outros. A melhor rotação é aquela em que os coeficientes do fator d_{ij} ou são muito pequenos (caso em que o fator associado é sem importância) ou muito grandes (quando o fator associado é muito importante). O tipo mais comum de rotação de fatores ortogonal é chamado **rotação varimax**, que maximiza o somatório da variância dos coeficientes dos fatores na Equação 12.28:

$$\text{var}_{\max} = \text{variância máxima}\left(\sum_{j=1}^{n} d_{ij}^2\right)$$

CALCULANDO E USANDO OS ESCORES DOS FATORES Os escores dos fatores para cada observação são calculados como os valores dos F_j para cada Y_i. Como os escores dos fatores originais F_j são combinações lineares dos dados (indo de trás para frente, da Equação 12.27 para Equação 12.25), podemos re-escrever a Equação 12.28 em notação de matrizes como:

$$\mathbf{F^* = (DTD)^{-1}DTY} \qquad (12.29)$$

onde **D** é a matriz $n \times n$ dos coeficientes dos fatores d_{ij}, **Y** é a de dados, e **F*** é a dos escores dos fatores rotacionados. Resolvendo esta equação diretamente, temos os escores dos fatores para cada réplica.

Tenha cuidado ao realizar testes de hipóteses com os escores dos fatores. Embora possam ser úteis para descrever dados multivariados, eles não são únicos (uma vez que existem infinitamente muitas escolhas de F_j), e o tipo de rotação usada é arbitrário. A análise fatorial é recomendada apenas para análise de dados exploratória.

UM BREVE EXEMPLO Conduzimos uma análise fatorial com os mesmos dados usados para a PCA: 10 variáveis padronizadas medidas em 39 plantas de *Darlingtonia* em quatro locais, em Oregon e na Califórnia (Tabela 12.3). A análise usou quatro fatores, que explicaram 28%, 26%, 13% e 8%, respectivamente, ou um total de 75% da variância nos dados.[11] Depois de usar a rotação varimax e resolver a Equação 12.29, plotamos os primeiros dois escores dos fatores para cada planta em um gráfico de dispersão (Figura 12.8). Como a PCA, a análise fatorial sugere alguma discriminação entre os grupos: as plantas do local HD (símbolos pretos) têm baixos escores no Fator 1, enquanto as do local TJH (símbolos verdes-claros) têm baixos escores no Fator 2.

Análise de coordenadas principais

A **análise de coordenadas principais** (*Principal Coordinates Analysis* – PCoA) é um método usado para ordenar dados usando qualquer medida de distância. A análise de componentes principais e a análise fatorial são empregadas quando usamos dados multivariados quantitativos, e desejamos preservar as distâncias euclidianas entre as observações. Em muitos casos, porém, as distâncias euclidianas entre as observações fazem pouco sentido. Por exemplo, uma matriz binária de presença e ausência é uma estrutura de dados muito comum em estudos ecológicos e ambientais: cada linha da matriz é um local ou uma amostra, e cada coluna é uma espécie ou um táxon. Os valores na matriz indicam a ausência (0) ou a presença (1) de uma espécie em um local. Como vimos anteriormente (Tabelas 12.4 e 12.5), distâncias euclidianas medidas para dados desse tipo podem dar resultados contraintuitivos. Outro exemplo no qual distâncias euclidianas não são apropriadas é uma matriz de distâncias genéticas das diferenças relativas em bandas de eletroforese. Em ambos, a PCoA é mais apropriada do que a PCA. A própria análise de componentes principais é um caso especial de PCoA: os

[11] Alguns programas de análise fatorial usam um teste da bondade do ajuste (Capítulo 11) para testar a hipótese de que o número de fatores usado na análise é suficiente para capturar o volume de variância contra a alternativa de que mais fatores são necessários. Para os dados de *Darlingtonia*, uma análise da bondade do ajuste de chi-quadrado falhou em rejeitar a hipótese nula de que quatro fatores eram suficientes ($\chi^2 = 17{,}91$, 11 graus de liberdade, $P = 0{,}084$).

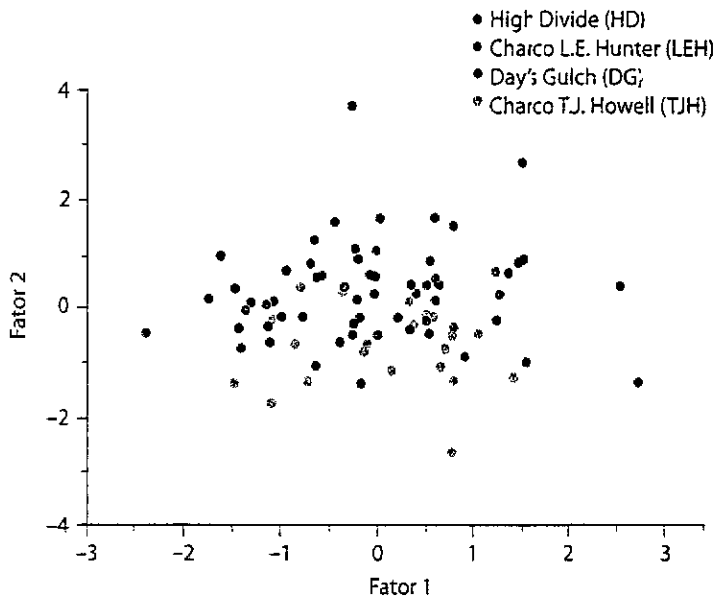


Figura 12.8 Gráfico ilustrando os primeiros dois fatores resultantes de uma análise fatorial dos dados da Tabela 12.3. Cada ponto representa uma planta individual e as cores representam os diferentes locais, como na Figura 12.6. Os resultados são qualitativamente similares à análise de componentes principais apresentada na Figura 12.6, com escores baixos para HD no Fator 1 e escores baixos para TJH no Fator 2.

autovalores e autovetores associados de uma PCA que calculamos de uma matriz de variância e covariância também podem ser recuperados aplicando-se PCoA a uma matriz de distâncias euclidianas.

CÁLCULOS BÁSICOS EM UMA ANÁLISE DE COORDENADAS PRINCIPAIS A análise de coordenadas principais segue cinco passos:

1. Gere uma matriz de distância ou dissimilaridade dos dados. Essa matriz, **D**, tem elementos d_{ij} que correspondem à distância entre duas amostras, i e j. Qualquer uma das medidas de distância dadas na Tabela 12.6 pode ser usada para calcular os d_{ij}. **D** é uma matriz quadrada, com o número de linhas i = ao número de colunas j = ao número de observações m.
2. Transforme **D** em uma nova matriz **D***, com elementos:

$$d^*_{i,j} = -\frac{1}{2} d^2_{i,j}$$

Essa transformação converte a matriz de distância em uma de coordenadas que preserva a relação de distância entre as variáveis transformadas e os dados originais.

TABELA 12.9 Matriz de presença e ausência usada para análise de coordenadas principais (PCoA), análise de correspondência (CA), e escalonamento multidimensional não métrico (NMDS)

	Amblyopone	*Aphaenogaster*	*Brachymyrmex*	*Camponotus*	*Crematogaster*	*Dolichoderus*	*Formica*
Connecticut	0	1	0	1	0	0	0
MA no continente	1	1	1	1	0	1	1
Ilhas de MA	1	0	0	0	1	1	1
Vermont	0	1	0	1	0	0	1

Cada linha representa amostras de diferentes localidades: Connecticut, Massachusetts (MA) no continente e em ilhas distantes de sua costa, e Vermont. Cada coluna representa um gênero diferente de formiga coletado durante um levantamento em florestas no entorno de pequenos brejos. (Dados compilados de Gotelli & Ellison, 2002a,b, e não publicados.)

3. Centralize a matriz **D*** para criar a Δ com elementos $\delta_{i,j}$, aplicando-se a seguinte transformação a todos os elementos de **D***:

$$\delta_{ij} = d^*_{ij} - \bar{d}^*_i - \bar{d}^*_j + \bar{d}^*$$

onde $\bar{d}^*_i$ é a média da linha i de **D***, $\bar{d}^*_i$ é a média da coluna j de **D***, e $\bar{d}^*_i$ é a média de todos os elementos de **D***.

4. Calcule os autovalores e autovetores de Δ. Os autovetores $\mathbf{a}_i$ devem ser colocados na escala da raiz quadrada dos seus autovalores correspondentes:

$$\sqrt{\mathbf{a}_k^T \mathbf{a}_k} = \sqrt{\lambda_k}$$

5. Escreva os autovetores $\mathbf{a}_k$ como colunas, com cada linha correspondendo a uma observação. Os valores são as novas coordenadas dos objetos no espaço de coordenadas principais, analogamente aos escores dos componentes principais.

UM BREVE EXEMPLO Usamos a PCoA para analisar uma matriz binária de presença e ausência de quatro locais (Connecticut, Vermont, Massachusetts no continente, e Ilhas de Massachusetts) e 16 gêneros de formigas encontradas em florestas em torno de pequenos brejos (Tabela 12.9). Primeiro, geramos uma matriz de distâncias entre os quatro locais usando a medida de dissimilaridade de Sørensen (Tabela 12.10). A matriz de distâncias 4×4 contém os valores das medidas de dissimilaridade de Sørensen calculadas para todas as combinações par-a-par de locais. Então, calculamos os autovetores, que são usados na contagem dos escores das coordenadas principais para cada local. Os escores das coordenadas principais para cada local estão plotados para os dois primeiros eixos de coordenadas principais (Figura 12.9). Os locais diferem conforme a dissimilaridade em suas composições de espécies e podem ser ordenados com base nisso. Os locais são

Lasius	*Leptothorax*	*Myrmecina*	*Myrmica*	*Paratrechina*	*Prenolepis*	*Ponera*	*Stenamma*	*Tapinoma*
1	1	0	1	0	0	0	1	1
1	1	1	1	0	1	0	1	0
1	1	1	1	1	1	0	1	1
1	1	0	1	0	0	0	1	1

ordenados CT, VT, MA no continente, ilhas de MA, da esquerda para a direita ao longo do primeiro eixo principal, que explica 80,3% da variância da matriz de distâncias. O segundo eixo principal separa, principalmente, os locais de MA no continente dos locais em ilhas de MA e explica uma variância adicional de 14,5% da matriz de distâncias.

Análise de correspondência

A **análise de correspondência** (*Correspondence Analysis* – CA), também conhecida como médias recíprocas (*Reciprocal Averaging* – RA) (Hill, 1973) ou análise de gradiente indireta, é usada para examinar a relação das assembleias de espécies às características dos locais. Os locais em geral são selecionados para abranger um gradiente ambiental, e a hipótese ou o modelo subjacente é que as distribuições das abundâncias das espécies são unimodais e aproximadamente normais (ou gaussianas) ao longo do gradiente ambiental (Whittaker, 1956). Os eixos resul-

TABELA 12.10 Medidas de dissimilaridade usadas na análise de coordenadas principais (PCoA) e no escalonamento multidimensional não métrico (NMDS)

	Connecticut	MA no continente	Ilhas de MA	Vermont
Connecticut	0	0,37	0,47	0,13
MA no continente	0,37	0	0,25	0,33
Ilhas de MA	0,47	0,25	0	0,43
Vermont	0,13	0,33	0,43	0

Os dados originais consistem de uma matriz de presença e ausência de gêneros de formigas de floresta em quatro locais na Nova Inglaterra (*ver* Tabela 12.9). Cada valor é a dissimilaridade de Sørensen entre cada par de locais (*ver* a fórmula na Tabela 12.6). Quanto mais dois locais diferem nos gêneros que eles contêm, maior é a dissimilaridade. Por definição, a matriz de dissimilaridade é simétrica e os elementos da diagonal iguais a 0,0. MA = Massachusetts.

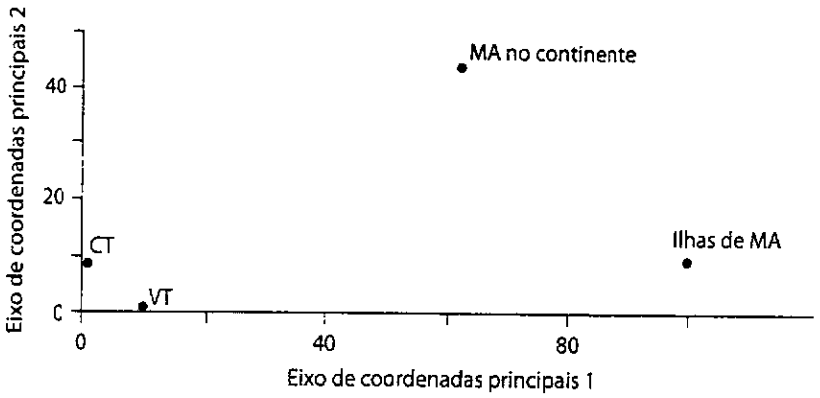


Figura 12.9 Ordenação de quatro locais de floresta com base nas suas assembleias de espécies de formigas (Tabela 12.9) usando Análise de Coordenadas Principais (PCoA). Uma matriz de distâncias par-a-par entre os quatro locais é calculada a partir desses dados (Tabela 12.10) e usada na PCoA. Os locais são ordenados CT, VT, MA no continente, ilhas de MA da esquerda para a direita ao longo do primeiro eixo principal, que explica 80,3% da variância da matriz de distâncias. O segundo eixo principal separa principalmente os locais MA no continente dos locais ilhas em MA, e explica uma variância adicional de 14,5%.

tantes de uma CA maximizam a separação das abundâncias das espécies ao longo de cada eixo de picos. A entrada de dados para uma análise de correspondência é uma matriz de locais × espécies que é manipulada como se fosse uma tabela de contingência (*ver* Capítulo 11). A análise de correspondência também pode ser considerada um caso especial de PCoA, e uma PCoA usando uma matriz de distância chi-quadrado resulta em uma CA.

CÁLCULOS BÁSICOS PARA UMA ANÁLISE DE CORRESPONDÊNCIA A análise de correspondência não é muito mais complexa de interpretar do que uma PCA ou análise fatorial, mas requer mais manipulações de matrizes.

1. Comece com uma tabela de linhas × colunas (como uma tabela de contingência ou uma matriz de locais [linhas] × espécies [colunas]) com m linhas e n colunas, e para a qual $m \geq n$. Esta frequentemente é uma restrição nos programas de CA. Note que se $m \leq n$ na matriz de dados originais, sua transposição atenderá a condição de que $m \geq n$, e a interpretação dos resultados será idêntica.
2. Crie uma matriz **Q** cujos elementos q_{ij} são proporcionais a valores de chi-quadrado:

$$q_{i,j} = \frac{\left(\dfrac{\text{observado} - \text{esperado}}{\sqrt{\text{esperado}}}\right)}{\sqrt{\text{total geral}}}$$

Os valores esperados são calculados como descrevemos no Capítulo 11 para testes de independência de chi-quadrado (Equação 11.4).

3. Aplique uma decomposição de valor singular a Q: $\mathbf{Q} = \mathbf{UWV}^T$, onde U é uma matriz $m \times n$, V é uma matriz $n \times n$, U e V são **matrizes ortonormais**, e W é uma matriz diagonal (*ver* Apêndice para detalhes).
4. Observe que o produto $\mathbf{Q}^T\mathbf{Q} = \mathbf{VW}^T\mathbf{WV}^T$.
5. A nova matriz diagonal $\mathbf{W}^T\mathbf{W}$, escrita como Λ, tem elementos λ_i que são autovalores de $\mathbf{Q}^T\mathbf{Q}$. A matriz V tem colunas que são autovetores nos quais os elementos são as cargas das colunas da matriz de dados originais. A matriz U tem colunas que são autovetores em que os elementos são as cargas das linhas da matriz de dados originais.
6. Use as matrizes U e V para plotar separadamente as localizações das linhas e das colunas no espaço de ordenação. As localizações também podem ser plotadas juntas em um ***bi-plot*** de ordenação, de modo que você possa ver as relações entre, digamos, os locais e as espécies. Entretanto, a análise de redundância (RDA; *ver* abaixo) é um método mais direto para verificar as relações conjuntas entre dados de locais (ambiental) e de espécies (composição).

Como a PCA e a análise fatorial, a CA resulta em eixos principais e escores, embora, neste caso, tenhamos escores tanto para as linhas (p. ex., locais) quanto para as colunas (p. ex., espécies). Assim como na PCA e na análise fatorial, o primeiro eixo de uma CA tem o maior autovalor (explica a maior parte da variância nos dados), bem como maximiza a associação entre linhas e colunas. Os eixos subsequentes explicam a variação residual e têm autovalores sucessivamente menores. Raras vezes usamos mais de dois ou três eixos de uma CA, porque imagina-se que cada eixo representa um gradiente ambiental particular, e porque um sistema com mais de três gradientes não relacionados é de difícil interpretação.

UM BREVE EXEMPLO Usamos a CA para examinar a relação conjunta da composição de gêneros de formiga em quatro locais: Connecticut, Massachusetts no continente e em ilhas distantes de sua costa e em Vermont (Tabela 12.9). Esses locais diferem ligeiramente na latitude ($\approx 3°$) e se estão no continente ou em ilhas.

Os primeiros dois eixos da CA explicam 58% e 28% (total = 86%) da variância dos dados, respectivamente. O gráfico da ordenação dos locais (Figura 12.10A) mostra uma discriminação entre os locais similar àquela observada com a PCoA: CT e VT se agrupam unidos, e são separados de MA ao longo do primeiro eixo. O segundo eixo principal separa os locais de MA no continente dos locais nas ilhas de MA. A separação dos gêneros com relação aos locais também é aparente (Figura 12.10B). Os gêneros com apenas uma ocorrência (como *Ponera* na amostra de VT, *Brachymyrmex* na amostra de MA no continente, e *Crematogaster* e *Paratrechina* na amostra de ilhas em MA) se separam da mesma forma que os

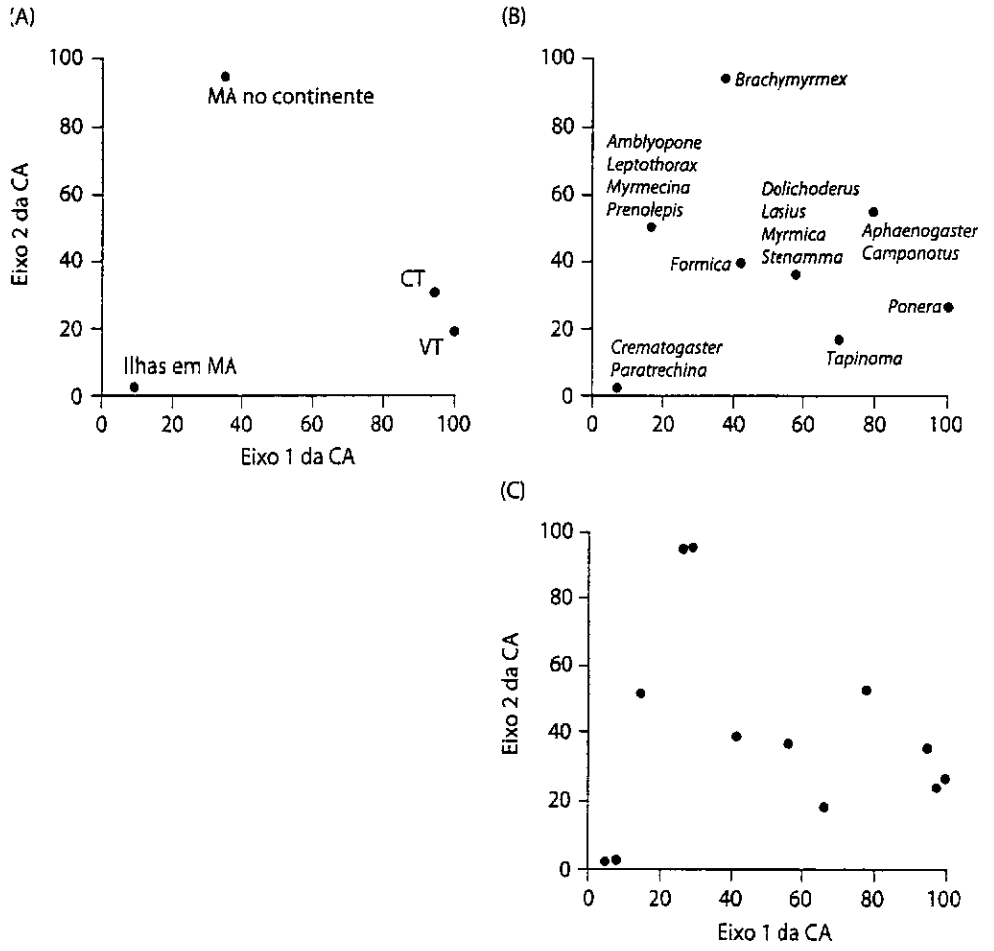


Figura 12.10 Resultados de uma análise de correspondência (CA) da matriz de locais × gêneros de formiga da Tabela 12.9. O gráfico A mostra a ordenação dos locais, o gráfico B, a ordenação dos gêneros, e o gráfico C é um *bi-plot* mostrando a relação entre as duas ordenações. Os resultados da ordenação dos locais são similares àqueles da análise de coordenadas principais (Figura 12.9), enquanto a ordenação dos gêneros separa-os com apenas uma ocorrência, como *Ponera*, na amostra de VT, *Brahymyrmex*, na amostra de MA no continente, e *Crematogaster* e *Paratrechina*, na de ilhas em MA. Os resultados da ordenação foram redimensionados de forma que as ordenações dos locais e das espécies possam ser apresentadas no mesmo gráfico. Em (C), os locais são indicados por pontos pretos e os gêneros de formigas por pontos verdes.

locais. Os gêneros remanescentes são arranjados mais para o centro do espaço de ordenação (Figura 12.10B). A sobreposição das duas ordenações captura as relações desses gêneros com os locais (compare a matriz de locais × gêneros na Tabela 12.9 com a Figura 12.10C). Usando mais informações sobre os locais (Gotelli & Ellison, 2002a), podemos tentar inferir informações adicionais sobre as relações entre cada gênero e as características dos locais.

Entretanto, você deve ser cauteloso ao fazer tais inferências. A dificuldade é que a análise de correspondência executa uma ordenação simultânea das linhas

e colunas da matriz de dados, e avalia quão bem as duas ordenações estão associadas uma com a outra. Assim, esperamos um grau de sobreposição razoável entre os resultados das duas ordenações (como na Figura 12.10C). Como os dados empregados na CA são essencialmente os mesmos daqueles usados na análise de tabela de contingência, o teste de hipótese e a modelagem preditiva são melhor realizados usando os métodos descritos no Capítulo 11.

O EFEITO DO ARCO E O DESTENDENCIAMENTO A análise de correspondência tem sido muito usada para explorar as relações entre as espécies ao longo de gradientes ambientais. Entretanto, como outros métodos de ordenação, ela tem a propriedade matemática lamentável de comprimir as extremidades de um gradiente ambiental e acentuar o centro. Isso pode resultar em gráficos de ordenação que são encurvados, como um arco ou em forma de ferradura (Legendre & Legendre, 1998; Podani & Miklós, 2002), mesmo quando as amostras são uniformemente espaçadas ao longo do gradiente ambiental. Um exemplo disso é visto na Figura 12.10A. O primeiro eixo dispõe os quatro locais linearmente; o segundo eixo puxa o local MA no continente para cima. Esses quatro pontos formam um arco relativamente suave. Em muitos casos, o segundo eixo da CA pode ser simplesmente uma distorção quadrática do primeiro eixo (Hill, 1973; Gauch, 1982).

Uma técnica de ordenação confiável e interpretável deve preservar as relações de distância entre os pontos, e seus valores originais e os criados pela ordenação devem ser igualmente distantes (até uma extensão possível, de acordo com o número de eixos que é selecionado) quando medidos. Para muitas medidas de distância comuns (incluindo a medida chi-quadrado, usada na CA), o **efeito do arco** distorce as relações de distância entre as novas variáveis criadas pela ordenação. A **análise de correspondência destendenciada** (*Detrended Correspondence Analysis* – DCA: Gauch, 1982) é usada para remover o efeito do arco da análise de correspondência e presumivelmente ilustrar mais acurácia na relação de interesse. Entretanto, Podani & Miklós (2002) mostraram que o arco é uma consequência matemática da aplicação da maioria das medidas de distância às espécies que respondem de forma unimodal aos gradientes ambientais subjacentes. Embora a DCA seja amplamente implementada em programas de ordenação, ela não é mais recomendada (Jackson & Somers, 1991). O uso de medidas de distância alternativas (Podani & Miklós, 2002) é uma solução melhor para o problema do arco.

Escalonamento multidimensional não métrico

Os quatro métodos de ordenação anteriores são similares no que diz respeito às distâncias entre as observações no espaço multivariado a serem preservadas, na medida do possível, depois dos dados multivariados terem sido reduzidos a um número menor de variáveis compostas. Em contraste, o objetivo no **escalonamento multidimensional não métrico** (*Non-Metric Multidimensional Scaling*

– NMDS) é terminar com um gráfico no qual objetos diferentes são posicionados distantes no espaço de ordenação, enquanto os similares são posicionados próximos. Apenas a ordem de classificação das distâncias ou dissimilaridades originais é preservada.

CÁLCULOS BÁSICOS DO ESCALONAMENTO MULTIDIMENSIONAL NÃO MÉTRICO Fazer um escalonamento multidimensional não métrico requer nove passos, alguns deles são repetidos.

1. Gere uma matriz de distância ou dissimilaridade **D** a partir dos dados. Qualquer medida de distância ou dissimilaridade pode ser usada (Tabela 12.6). Os elementos d_{ij} em **D** são distâncias ou dissimilaridades entre as observações.
2. Escolha o número de dimensões (eixos) n para ser usado ao fazer a ordenação. Use dois ou três eixos, porque a maioria dos gráficos é bi (eixos x e y) ou tridimensional (eixos x, y e z).
3. Comece a ordenação posicionando as m observações no espaço n dimensional. As análises subsequentes dependem muito desta inicialização, pois o NMDS encontra sua solução por minimização local (similar à regressão não linear; *ver* Capítulo 9). Se alguma informação geográfica estiver disponível (como latitude e longitude), pode ser usada como ponto de partida. Alternativamente, o resultado de outra ordenação (como a PCoA) pode ser usado para determinar a posição inicial das observações em um NMDS.
4. Calcule novas distâncias δ_{ij} entre as observações na configuração inicial. Normalmente, distâncias euclidianas são usadas para calcular δ_{ij}.
5. Faça uma regressão de δ_{ij} contra d_{ij}. O resultado dessa regressão é um conjunto de valores preditos $\hat{\delta}_{ij}$. Por exemplo, se usarmos uma regressão linear, $\delta_{ij} = \beta_0 + \beta_1 d_{ij} + \varepsilon_{ij}$, então os valores preditos são:

$$\hat{\delta}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 d_{ij}$$

6. Calcule a bondade do ajuste entre δ_{ij} e $\hat{\delta}_{ij}$. A maioria dos programas de NMDS calcula como um **estresse**:

$$\text{Estresse} = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(\delta_{ij} - \hat{\delta}_{ij})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n}\hat{\delta}_{ij}^2}}$$

O estresse é calculado sobre o triângulo inferior da matriz quadrada **D**. O numerador é a soma geral do quadrado da diferença entre os valores observados e esperados, e o denominador é o grande somatório dos quadrados dos valores esperados. Note que o estresse parece muito com uma estatística de teste chi-quadrado (*ver* Capítulo 11).

7. Mude levemente a posição das m observações no espaço n dimensional para reduzir o estresse.
8. Repita os passos 4 a 7 até que o estresse não possa mais ser reduzido.
9. Plote a posição das m observações no espaço n dimensional para que o estresse seja mínimo. Este gráfico ilustra a "relação" entre as observações.

UM BREVE EXEMPLO Usamos NMDS para analisar os dados de formigas na Tabela 12.8. Adotamos a medida de dissimilaridade de Sørensen como nossa medida de distância. O *scree plot* dos escores de estresse indica que duas dimensões são suficientes para a análise (Figura 12.11). Como na CA e na PCoA, temos uma boa discriminação entre os locais (Figura 12.12A), e boa separação entre assembleias de gêneros com apenas uma ocorrência (Figura 12.12B). Como a escala dos eixos no NMDS é arbitrária, não podemos sobrepor esses dois gráficos em um *bi-plot* de ordenação.

Vantagens e desvantagens da ordenação

Os ecólogos e cientistas da área ambiental usam a ordenação para: reduzir dados multivariados complexos para um conjunto de dados menor e mais manejável, para ordenar os locais com base nas variáveis ambientais de assembleias de espécies e para identificar respostas das espécies aos gradientes ou perturbações ambientais. Como alguns métodos de ordenação (como a PCA e a análise fatorial) em geral estão disponíveis na maioria dos pacotes estatísticos comerciais, temos acesso fácil, e frequentemente com menus, a essas ferramentas. Quando usada

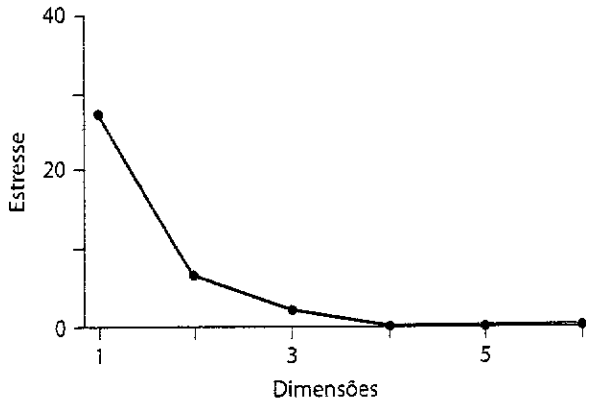


Figura 12.11 *Scree plot* para as dimensões em uma análise de Escalonamento Multidimensional Não Métrico (NMDS) dos dados de locais × gêneros de formiga (Tabela 12.9). Uma matriz de distâncias par-a-par para os quatro locais é calculada a partir desses dados (Tabela 12.10) e usada no NMDS. Em um NMDS, o estresse nos dados é uma medida de desvio similar a uma estatística de bondade do ajuste de chi-quadrado. Esse *scree plot* ilustra a redução sucessiva no estresse com o aumento de dimensões no NMDS dos dados de formigas. Nenhuma redução significativa no estresse é obtida com mais de três dimensões, explicando a maior parte da bondade do ajuste.

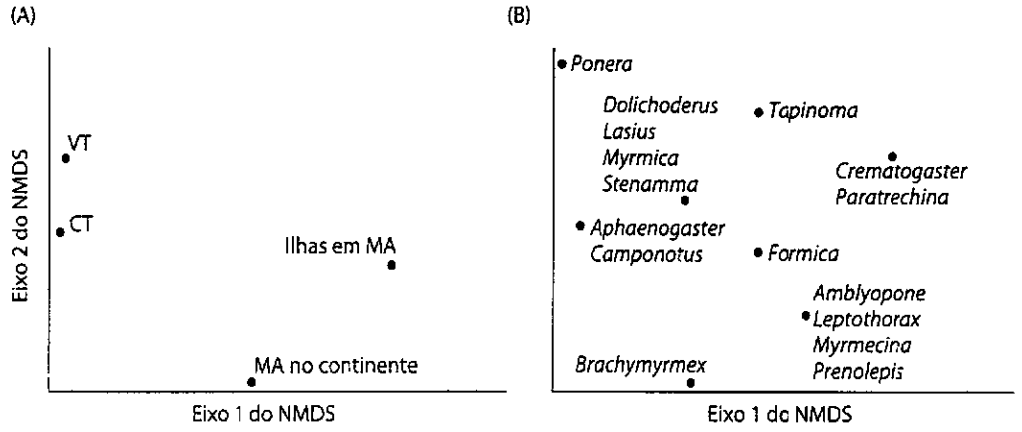


Figura 12.12 Gráfico cas duas primeiras dimensões da análise de Escalonamento Multidimensional Não Métrico (NMDS) dos dados de locais × gêneros de formiga da Tabela 12.9. Uma matriz de distâncias par-a-par para os quatro locais é calculada a partir desses dados (Tabela 12.10) e usada no NMDS. (A) Ordenação dos locais. (B) Ordenação dos gêneros.

com prudência, a ordenação pode ser uma ferramenta poderosa na exploração de dados, para ilustrar padrões e gerar hipóteses que possam ser testadas com amostragens ou experimentos subsequentes.

As desvantagens da ordenação não são aparentes, porque essas técnicas têm sido automatizadas em vários programas, os mecanismos são ocultos e os pressupostos raras vezes são enunciados. A ordenação baseia-se em álgebra de matrizes, com a qual muitos ecólogos e cientistas da área ambiental não são familiares. Como os eixos principais em geral são redimensionados antes de plotar os escores principais, estes são interpretáveis apenas em relação uns aos outros, sendo difícil relacioná-los de volta às medidas originais.

RECOMENDAÇÕES Raras vezes fica evidente qual técnica de ordenação você deve escolher no intuito de melhor ordenar as observações, amostras ou populações no espaço multivariado. A simples redução dos dados é melhor efetuada pela análise de componentes principais (PCA) quando distâncias euclidianas são apropriadas e não há presença de dados discrepantes ou dados muito assimétricos. A análise de coordenadas principais (PCoA) é apropriada quando outras medidas de distância são necessárias (a PCA é equivalente a uma PCoA quando distâncias euclidianas são empregadas). Recomendamos a PCoA para a maioria das aplicações de ordenação nas quais o objetivo é preservar as distâncias multivariadas originais entre as observações no espaço reduzido (ordenação). O escalonamento multidimensional não métrico (NMDS) preserva apenas a ordem de classificação das distâncias, mas pode ser usado com qualquer medida de distância. A análise de correspondência (CA) é um método de ordenação razoável para examinar a distribuição de espécies ao longo de gradientes ambientais, mas a sua prima, a análise de correspondência destendenciada (DCA),

não deve mais ser usada. Da mesma forma, os resultados da análise fatorial são difíceis de interpretar e muito dependentes de métodos de rotação subjetivos. Em todos os casos, é importante lembrar as questões ecológicas e ambientais de interesse subjacentes. Por exemplo, se a questão original era ver como espécies ou características são distribuídas ao longo de um gradiente ambiental, então simplesmente plotar a variável resposta contra uma medida apropriada do ambiente (ou o primeiro ou o segundo eixo principal) pode ser mais informativo que qualquer gráfico de ordenação.

Independentemente do método, a ordenação deve ser usada com cautela para o teste de hipóteses, sendo mais útil para a exploração de dados e geração de padrões. Entretanto, os testes de hipóteses clássicos podem ser realizados sobre os escores resultantes de qualquer ordenação, desde que os pressupostos dos testes sejam atendidos.

CLASSIFICAÇÃO

A **classificação** é o processo através do qual agrupamos os objetos. Enquanto o objetivo da ordenação é separar as observações ou amostras ao longo de gradientes ambientais ou eixos biológicos, o objetivo da classificação é agrupar objetos similares em classes identificáveis e interpretáveis que podem ser distinguidas das classes vizinhas. A taxonomia é uma aplicação familiar da classificação – um taxonomista começa com uma coleção de espécimes e deve decidir como atribuí-los em espécies, gêneros, famílias e grupos taxonômicos superiores. Independente de a taxonomia ser fundamentada na morfologia, em sequências gênicas, ou na história evolutiva, o objetivo é o mesmo: um sistema de classificação inclusiva com base em grupos de agrupamentos hierárquicos.

Os ecólogos e cientistas da área ambiental podem desconfiar de classificações porque se assume que as classes representam entidades discretas, enquanto estamos mais acostumados a lidar com variação contínua na estrutura de comunidades ou na morfologia que está mapeada em gradientes ambientais contínuos. A ordenação identifica padrões dentro de gradientes, enquanto a classificação identifica pontos finais ou extremos dos gradientes enquanto ignora os intermediários.

Análise de agrupamentos

O tipo mais familiar de análise de classificação é a **análise de agrupamentos** (*cluster analysis*). A análise de agrupamentos pega m observações, das quais cada uma está associada com n variáveis numéricas contínuas e segrega as observações em grupos. A Tabela 12.11 ilustra o tipo de dados usados na análise de agrupamentos. Cada uma das nove linhas da tabela representa um país diferente onde as amostras foram coletadas. Cada uma das três colunas da tabela representa uma medida morfológica obtida de espécimes do caramujo marinho *Littoraria angulifera*. O

TABELA 12.11 Medidas de conchas de caramujos usadas na análise de agrupamentos

País	Continente	Proporcionalidade	Circularidade	Altura do espiral
Angola	África	1,36	0,76	1,69
Bahamas	América do N.	1,51	0,76	1,86
Belize	América do N.	1,42	0,76	1,85
Brasil	América do S.	1,43	0,74	1,71
Flórida	América do N.	1,45	0,74	1,86
Haiti	América do N.	1,49	0,76	1,89
Libéria	África	1,36	0,75	1,69
Nicarágua	América do N.	1,48	0,74	1,69
Serra Leoa	África	1,35	0,73	1,72

A forma da concha do caramujo *Littoraria angulifera* foi medida para amostras de 9 países. Proporcionalidade da concha = altura/largura da concha, circularidade da concha = largura/altura da abertura e altura do espiral = altura da concha/comprimento da abertura. Os valores na tabela são médias com base em 2 a 100 amostras por local (dados de Merkt & Ellison, 1998). A análise de agrupamentos reúne os diferentes países com base na similaridade na morfologia da concha (*ver* Figura 12.13).

objetivo dessa análise é formar agrupamentos de locais com base na similaridade da forma da concha. A análise de agrupamentos também pode ser usada para agrupar locais com base nas abundâncias, ou presença e ausência de espécies, ou para agrupar organismos conforme a similaridade de características medidas, como a morfologia ou as sequências de DNA.

Escolhendo um método de agrupamento

Existem vários métodos disponíveis para agrupar dados (Sneath & Sokal, 1973), focalizaremos em apenas dois, os quais frequentemente são usados por ecólogos e cientistas da área ambiental.

AGRUPAMENTO AGLOMERATIVO *VERSUS* DIVISIVO O **agrupamento aglomerativo** procede tomando várias observações separadas e as reunindo em agrupamentos sucessivamente maiores até que apenas um seja obtido. O **agrupamento divisivo**, por outro lado, procede colocando todas as observações em um grupo e depois dividindo sucessivamente, até que cada observação esteja em seu próprio agrupamento. Em ambos os métodos, uma decisão deve ser tomada, muitas vezes com base em alguma regra estatística, sobre quantos agrupamentos usar para descrever os dados.

Ilustramos esses dois métodos usando os dados na Tabela 12.11. Os métodos aglomerativos começam com uma matriz quadrada $m \times m$ de distâncias, na qual os valores são as distâncias morfológicas par-a-par medidas para cada par de locais; as distâncias euclidianas são as mais simples de usar (Tabela 12.12),

TABELA 12.12 Matriz de distância euclidiana para as medidas morfológicas das conchas de caramujos

	Angola	Bahamas	Belize	Brasil	Flórida	Haiti	Libéria	Nicarágua	Serra Leoa
Angola	0								
Bahamas	0,23	0							
Belize	0,17	0,09	0						
Brasil	0,08	0,17	0,14	0					
Flórida	0,19	0,06	0,04	0,15	0				
Haiti	0,24	0,04	0,08	0,19	0,05	0			
Libéria	0,01	0,23	0,17	0,07	0,19	0,24	0		
Nicarágua	0,12	0,17	0,17	0,05	0,17	0,20	0,12	0	
Serra Leoa	0,04	0,21	0,15	0,08	0,17	0,22	0,04	0,13	0

Os dados originais estão detalhados na Tabela 12.11. Como as matrizes de distância são simétricas, apenas a metade inferior é apresentada.

mas qualquer medida de distância (Tabela 12.6) pode ser empregada na análise de agrupamentos. Uma análise do tipo aglomerativa começa com todos os objetos – locais, nesta análise – em separado e sucessivamente os agrupa. Vimos, na matriz de distância euclidiana (Tabela 12.12), que o Brasil é mais próximo (em unidades de distância) da Nicarágua, a Angola é mais próxima da Libéria, Belize é mais próximo da Flórida e as Bahamas são mais próximas do Haiti. Esses agrupamentos formam a linha de baixo do **dendrograma** (diagrama de árvore) mostrado na Figura 12.13. Neste exemplo, Serra Leoa é o local ímpar (sem par, pois o agrupamento aglomerativo trabalha para cima e em pares) e é adicionada ao grupo Angola-Libéria para formar um novo agrupamento africano. Os quatro novos grupos resultantes são reunidos mais adiante, de tal maneira que os agrupamentos sucessivamente maiores contêm os menores. Desse modo, o agrupamento Bahamas-Haiti é ligado ao grupo Belize-Flórida para formar um novo agrupamento Caribe-Atlântico Norte; os outros dois grupos formam o Atlântico Sul-África.

Uma análise de agrupamentos divisiva começa com todos os objetos em um grupo, e em seguida os divide com base na dissimilaridade. Neste caso, os grupos resultantes são os mesmos que os produzidos no agrupamento aglomerativo. Mais tipicamente, o agrupamento divisivo resulta em menos grupos, cada um com mais objetos, enquanto o aglomerativo resulta em mais grupos, cada um com menos objetos. Os algoritmos de agrupamento divisivo também podem avançar até um número pré-determinado de grupos, que é especificado pelo investigador antes de realizar a análise. Na atualidade, um método raras vezes usado de classificação de comunidades, o TWINSPAN (sigla em inglês para "análise de espécies indicadoras com dois fatores"; *ver* Gauch, 1982), usa algoritmos de agrupamento divisivo.

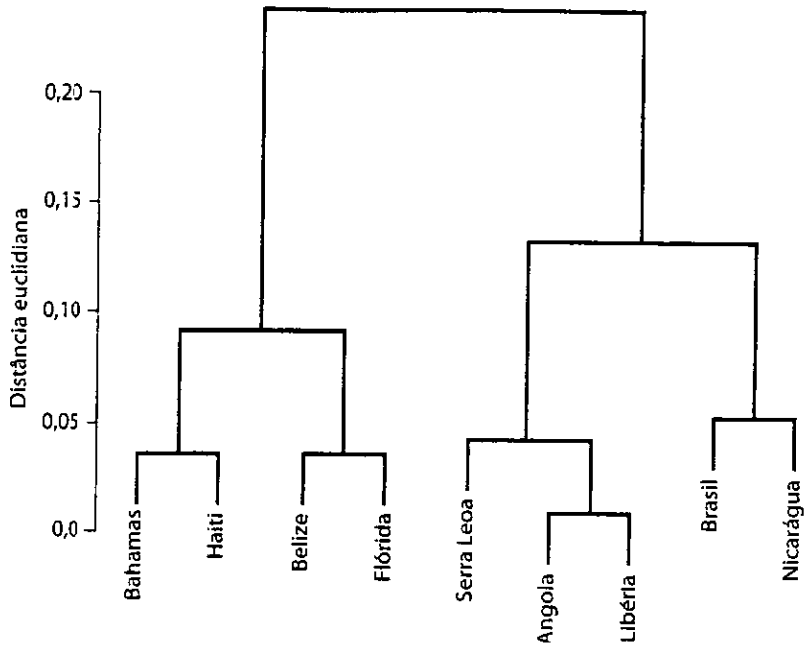


Figura 12.13 Resultados de uma análise de agrupamentos aglomerativos. Os dados originais (Merkt & Ellison, 1998) consistem em taxas morfológicas de medidas da concha de caramujos (*Littoraria angulifera*) de vários países (Tabela 12.11). Um dendrograma (figura ramificada) reúne os locais em grupos com base na similaridade na morfologia da concha. O eixo-*y* indica a distância euclidiana na qual os locais (os grupos de níveis inferiores) são ligados aos grupos de níveis superiores.

MÉTODOS HIERÁRQUICOS *VERSUS* NÃO HIERÁRQUICOS Intuitivamente, grupos com poucas observações deveriam ser incorporados naqueles de ordem superior, com mais observações. Em outras palavras, se as observações *a* e *b* estão em um grupo, e as observações *c* e *d* estão em outro, um grupo maior, capaz de incluir *a* e *c*, também deve adotar *b* e *d*. Esse arranjo de classificação deve ser familiar da taxonomia: se duas espécies estão em um gênero e duas outras estão em outro gênero, e os dois gêneros estão na mesma família, então todas as quatro espécies também estão na mesma família. Aqueles que seguem essa regra são chamados **métodos de agrupamento hierárquicos**. Tanto os aglomerativos quanto os divisivos podem ser hierárquicos.

Em contraste, os métodos de **agrupamento não hierárquicos** reúnem as observações sem depender de um sistema de referência externamente imposto. A ordenação pode ser considerada análoga ao agrupamento não hierárquico – os resultados de uma ordenação com frequência são independentes de ordenações do sistema externamente impostas. Por exemplo, duas espécies do mesmo gênero podem terminar em diferentes grupos produzidos pela ordenação de um conjunto de características morfológicas ou do hábitat. No passado, os ecólogos e cientistas da área ambiental eram mais propensos a usar métodos não hierárquicos do

que os hierárquicos, visto que o agrupamento era usado para buscar padrões nos dados e para gerar hipóteses. Contudo, conforme mais estudos começaram a incorporar dados filogenéticos (Losos, 1996), os métodos hierárquicos puderam ser mais usados, pois as filogenias conhecidas impõem uma estrutura de referência externa ao sistema de estudo.

O **agrupamento por *K*-médias** (*K-means*) é um método não hierárquico que, primeiro, requer que você especifique quantos grupos deseja. O algoritmo cria grupos de tal forma que todos os objetos internos estejam mais próximos uns dos outros (com base em uma medida de distância) do que eles estão de objetos dentro de outros grupos. O agrupamento por *K*-médias minimiza o somatório dos quadrados das distâncias de cada objeto ao centroide do seu grupo.

Primeiro, aplicamos o agrupamento por *K*-médias aos dados de caramujos especificando dois grupos. Um, consistindo de amostras da Angola, Brasil, Libéria, Nicarágua e Serra Leoa, com conchas relativamente pequenas (centroide: proporcionalidade = 1,39; circularidade = 0,74; altura do espiral = 1,70), o outro (Bahamas, Belize, Flórida e Haiti) contendo conchas relativamente grandes (centróide = 1,47; 0,76; 1,87). O somatório interno dos quadrados para o primeiro grupo (África-Atlântico Sul) é igual a 0,14, e, para o segundo grupo (Caribe-Atlântico Norte), 0,06. Um agrupamento por *K*-médias desses dados especificando três grupos divide o grupo África-Atlântico Sul em dois, um para os locais africanos e outro para a Nicarágua e o Brasil.

Como escolher quantos grupos usar? Não existe nenhuma regra rígida e rápida, mas quanto mais grupos se usa menor é o número de membros em cada um, e mais similares os grupos serão uns com os outros. Hartigan (1975) sugere que primeiro você faça um agrupamento por *K*-médias com k e $k + 1$ grupos em uma amostra de m observações. Se:

$$\left[\frac{\sum SS_{(dentro\ de\ K\ grupos)}}{\sum SS_{(dentro\ de\ K+1\ grupos)}} \times (m-k-1)\right] > 10 \qquad \textbf{(12.30)}$$

então faz sentido adicionar o $(k + 1)$-ésimo grupo. Se a desigualdade na Equação 12.30 for inferior a 10, é melhor parar com k grupos. A Equação 12.30 irá gerar um número grande sempre que a adição de um grupo aumente substancialmente a soma dos quadrados dentro de grupos. No entanto, em estudos de simulação, a Equação 12.30 tende a superestimar o número real de grupos (Sugar & James, 2003). Para os dados de caramujos da Tabela 12.11, a Equação 12.30 = 8,1, sugerindo que dois grupos são melhores que três. Sugar & James (2003) revisam uma ampla gama de métodos para identificar o número de grupos em um conjunto de dados, uma área ativa de pesquisa em análise multivariada.

Análise discriminante

A **análise discriminante** é usada para atribuir amostras a grupos ou classes que são definidas *a priori*; ela se comporta como uma análise de agrupamentos ao

avesso. Retomando o exemplo dos caramujos, podemos usar as localizações geográficas para definir grupos *a priori*. As amostras de caramujos vêm de três continentes: África (Angola, Libéria e Serra Leoa), América do Sul (Brasil), e América do Norte (todo o resto). Se nos tivessem passado apenas os dados de forma da concha (três variáveis), poderíamos atribuir as amostras ao continente correto? A análise discriminante é usada para gerar combinações lineares de variáveis (como na PCA: Equação 12.21) que são mais tarde usadas para separar o melhor possível dos grupos. Como na MANOVA, a análise discriminante requer que os dados se conformem a uma distribuição normal multivariada.

Podemos gerar uma combinação linear das variáveis originais, usando a Equação 12.21:

$$Z_i = a_{i1}Y_1 + a_{i2}Y_2 + \ldots + a_{in}Y_n$$

Se as médias multivariadas $\overline{\mathbf{Y}}$ variam entre os grupos, queremos encontrar os coeficientes a da Equação 12.20 que maximizam as diferenças entre os grupos. Especificamente, os coeficientes a_{in} são aqueles que *maximizam* a razão-F em uma ANOVA – em outras palavras, queremos que a soma dos quadrados entre grupos seja a *maior* possível para um dado conjunto de Z_i's.

Mais uma vez, retornamos às nossas matrizes **H** (entre grupos) e **E** (dentro de grupos, ou erro; *ver* a discussão anterior, na MANOVA) da soma dos quadrados e dos produtos cruzados (SQPC). Os autovalores e autovetores da matriz $\mathbf{E}^{-1}\mathbf{H}$ são então determinados. Se ordenarmos os autovalores do maior para o menor (como na PCA, por exemplo), seus autovetores $\mathbf{a}_i$ correspondentes são os coeficientes da Equação 12.21.

Os resultados desta análise discriminante são dados na Tabela 12.13. Este pequeno subconjunto de todo o conjunto de dados se desvia da normalidade multivariada ($E_n = 31{,}32$, $P = 2 \times 10^{-5}$, com 6 graus de liberdade), mas o conjunto de dados completo (de 1.042 observações) passou nesse teste de normalidade multivariada ($P = 0{,}13$). Embora as médias multivariadas (Tabela 12.13A) não difiram significativamente por continente (lambda de Wilk = 0,108, F = 2,7115, $P = 0{,}09$ com 6 e 8 graus de liberdade), ainda podemos ilustrar a análise discriminante.[12] Os primeiros dois autovetores (Tabela 12.13B) explicaram 97% da variância nos dados, e foram usados na Equação 12.21 para calcular os escores dos componentes principais para cada uma das amostras originais. Os escores das nove observações da Tabela 12.11 estão plotados na Figura 12.14, com cores diferentes, correspondendo a conchas de diferentes continentes. Vemos um agrupamento óbvio das amostras africanas, e outro agrupamento evidente de amostras da América do Norte. Entretanto, uma das amostras da América do Norte está mais próxima da América do Sul.

[12] O conjunto de dados completo, analisado por Merkt & Ellison (1998), tinha 1.042 amostras de 19 países. A análise discriminante foi usada para predizer de qual sistema de correntes oceânicas (caribenha, do Golfo do México, corrente do Golfo, Equatorial Norte ou Sul) uma concha veio. A análise discriminante atribuiu com sucesso mais de 89% das amostras ao sistema de correntes correto.

TABELA 12.13 Análise discriminante por continente dos dados de morfologia da concha de caramujos em 9 locais

A. Médias das variáveis

	África	América do N.	América do S.
Proporcionalidade	1,357	1,470	1,430
Circularidade	0,747	0,752	0,740
Altura do espiral	1,700	1,830	1,710

O primeiro passo é calcular a média variável para as amostras agrupadas por continente.

B. Autovetores

	a_1	a_2
Proporcionalidade	0,921	0,382
Circularidade	–0,257	–0,529
Altura do espiral	0,585	–0,620

A análise discriminante gera combinações lineares dessas variáveis (similar à PCA), que maximiza a separação das amostras por continente. Os autovetores contêm os coeficientes (cargas), que são usados para criar um escore discriminante para cada amostra, apenas os dois primeiros, que explicam 97% da variância, são mostrados.

C. Matriz de classificação

		Predito (Classificado)			
		África	América do N.	América do S.	%Corretos
Observado	África	3	0	0	100
	América do N.	0	4	1	80
	América do S.	0	0	1	100

As amostras são atribuídas a um dos três continentes de acordo com seus escores discriminantes. Nessa matriz de classificação, todas as amostras, exceto uma, foram corretamente atribuídas ao seu continente original. Esse resultado não é muito surpreendente, porque os mesmos dados foram usados para criar o escore discriminante e para classificar as amostras.

D. Matriz de classificação por Jackknife

		Predito (Classificado)			
		África	América do N.	América do S.	%Corretos
Observado	África	2	0	1	67
	América do N.	0	3	2	60
	América do S.	0	1	0	0

Uma abordagem menos enviesada é usar uma matriz de classificação por jackknife, na qual o escore discriminante é criado a partir de $n - 1$ amostras e depois usado para classificar a amostra excluída. Desta forma, a função discriminante é independente da amostra que está sendo classificada. A classificação por jackknife não teve uma performance tão boa quanto as outras matrizes, com apenas 5 das 9 amostras corretamente classificadas por continente.

O conjunto de dados completo do qual essas relações foram retiradas é detalhado na Tabela 12.11. A análise discriminante e outras técnicas multivariadas devem ser fundamentadas em amostras muito maiores que as usadas neste exemplo. (Dados de Merkt & Ellison, 1998.)

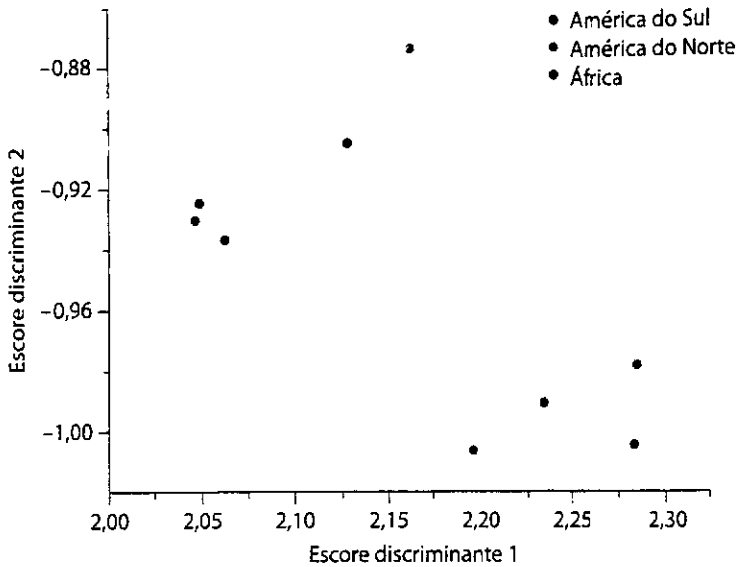


Figura 12.14 Um gráfico dos dois primeiros eixos de uma análise discriminante de medidas de razões morfológicas em conchas de caramujos (*Littoraria angulifera*) de vários países (Tabela 12.11). A análise discriminante é usada para avaliar quão bem as conchas dos três continentes diferentes podem ser distinguidas. Os escores discriminantes foram calculados usando os autovetores da Tabela 12.13 e a Equação 12.21.

As classificações preditas baseiam-se na minimização das distâncias (geralmente a de Mahalanobis ou euclidiana; *ver* Tabela 12.6) entre as observações dentro de grupos. A aplicação desse método aos dados resulta em predições acuradas para as conchas sul-americanas e africanas, mas apenas 80% (4 de 5) das conchas norte-americanas foram corretamente classificadas. A Tabela 12.13C resume essas classificações na forma de **matriz de classificação**. As linhas da matriz de classificação quadrada representam os grupos observados e as colunas, os grupos preditos (classificados). Os valores representam o número de observações que foram classificadas em um grupo particular. Se o algoritmo de classificação não tiver erros, então todas as observações cairão na diagonal. Os valores fora da diagonal representam erros na classificação em que uma observação foi atribuída de forma incorreta.

Não deve ser nenhuma surpresa que a classificação predita em uma análise discriminante corresponda aos dados observados, afinal, esses dados foram usados para gerar os autovetores, que depois usamos para classificar os dados. Assim, esperamos que os resultados sejam enviesados em favor da correta atribuição das nossas observações a grupos dos quais elas vieram. Uma solução para superar esses vieses é a randomização por jackknife da matriz de classificação da Tabela 12.13C. O jackknife retira uma observação individual do conjunto de dados, reanalisa os dados sem aquela observação, depois usa os resultados para classificar a observação retirada (*ver* também "A Função de Influência" no Capítulo 9, e a Nota

de rodapé 2, Capítulo 5). A matriz de classificação de jackknife é dada na Tabela 12.13D. Essa classificação é muito pobre, pois as médias multivariadas para cada continente não foram muito diferentes. Por outro lado, normalmente não faríamos uma análise discriminante em um conjunto de dados tão pequeno. Outra solução, se o conjunto de dados original for grande, seria dividi-lo de maneira aleatória em dois grupos: um para construir a classificação e o outro para testá-la. Novos dados também podem ser coletados e classificados com a função discriminante, embora precise haver um jeito de verificar independentemente que a atribuição dos grupos está correta.

Vantagens e desvantagens da classificação

As técnicas de classificação são relativamente fáceis de usar e interpretar, e estão amplamente implementadas nos programas de estatística. Elas nos permitem tanto agrupar amostras, quanto atribui-las a grupos identificados *a priori*. Os dendrogramas e as matrizes de classificação resumem e transmitem claramente os resultados da análise de agrupamentos e da discriminante.

No entanto, tanto a análise de agrupamentos quanto a discriminante devem ser usadas com cautela. Existem diversas maneiras de fazer uma análise de agrupamentos, e os diferentes métodos em geral dão resultados diferentes. É mais importante decidir um método de antemão em vez de tentar todos os diferentes métodos e escolher aquele que parecer o mais agradável ou que se conforme a noções preconcebidas. A análise discriminante é muito mais simples de fazer, porém ela requer dados extensos e a atribuição de grupos *a priori*. As análises de agrupamentos e discriminante são métodos descritivos e exploratórios. A estatística da MANOVA pode ser usada com os resultados de uma análise discriminante para testar hipóteses acerca das diferenças entre grupos. Inversamente, a análise discriminante também pode ser usada como um teste *a posteriori* para comparar grupos, uma vez que a hipótese nula em uma MANOVA tenha sido rejeitada.

De fato, as análises de agrupamentos e discriminante partem do pressuposto da existência de grupos ou agrupamentos distintos, enquanto uma hipótese nula adequada seria de que as amostras foram obtidas de um único grupo e que exibem apenas diferenças aleatórias entre elas. Strauss (1982) descreve métodos de Morte Carlo para testar a significância estatística de agrupamentos. O *bootstrap* e outros métodos computacionalmente intensivos para testar a significância estatística de grupos também são muito utilizados em análises filogenéticas (Felsenstein, 1985; Emerson et al., 2001; Huelsenbeck et al., 2002; Sanderson & Shaffer, 2002; Miller, 2003; Sanderson & Driskell, 2003).

REGRESSÃO MÚLTIPLA MULTIVARIADA

Os métodos para análise multivariada descritos até agora foram primariamente descritivos (ordenação, classificação), ou limitados a variáveis preditoras cate-

góricas (teste T^2 de Hotelling e MANOVA). O tópico final deste capítulo é uma extensão da regressão (uma variável resposta contínua e uma ou mais variáveis preditoras contínuas) para dados de resposta multivariados. Dois métodos são comumente usados por ecólogos: a **análise de redundância** (*Redundancy Analysis* – **RDA**) e a **análise de correspondência canônica** (*Canonical Correspondence Analysis* – **CCA**). Descrevemos apenas os mecanismos da RDA, que é uma extensão direta da regressão múltipla para o caso multivariado. A RDA assume uma relação causal entre as variáveis dependentes e independentes, enquanto a CCA tem como foco a geração de um eixo unimodal em relação às variáveis respostas (p. ex., ocorrências ou abundâncias de espécies) e um eixo linear em relação às variáveis preditoras (p. ex., características do hábitat ou do ambiente). Uma terceira técnica, a análise de correlação canônica (Hotelling, 1936; Manly, 1991) baseia-se nas relações simétricas entre as duas variáveis. Em outras palavras, a análise de correlação canônica assume erro tanto nos preditores quanto nas respostas (*ver* discussão dos pressupostos da regressão no Capítulo 9).

Análise de redundância

Um dos usos mais comuns da RDA é para examinar as relações entre composição de espécies, medida como as abundâncias de cada uma de n espécies, e características ambientais, medida como um conjunto de m variáveis ambientais (Legendre & Legendre, 1998). Os dados de composição de espécies representam a variável resposta multivariada, e as variáveis ambientais representam a variável preditora multivariada. Todavia, a RDA não é restrita a este tipo de análise – ela pode ser usada para qualquer regressão múltipla envolvendo múltiplas variáveis respostas e múltiplas variáveis preditoras, como demonstraremos em nosso exemplo.

O BÁSICO EM UMA ANÁLISE DE REDUNDÂNCIA Como desenvolvido no Capítulo 9, a regressão múltipla linear padrão é:

$$Y_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon_j \qquad (12.31)$$

Os valores de β_i são os parâmetros a serem estimados, e ε_j representa o erro aleatório. Os métodos padrão de mínimos quadrados são usados para ajustar o modelo e obter estimativas não enviesadas de $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\varepsilon}}$ dos β_i e dos ε_j. No caso multivariado, a única variável resposta Y é substituída por uma matriz $\mathbf{Y}$ com n variáveis. A análise de redundância faz a regressão de $\mathbf{Y}$ contra uma matriz de variáveis independentes $\mathbf{X}$, todas medidas para cada observação. Os passos de uma RDA são:

1. Faça uma regressão de cada uma das variáveis respostas individuais Y_j de $\mathbf{Y}$ contra todas as variáveis de $\mathbf{X}$ (usando a Equação 12.31). Isso resulta em uma

matriz de valores ajustados $\hat{\mathbf{Y}} = \left[\hat{Y}_j\right]$. Essa matriz é calculada como $\hat{\mathbf{Y}} = \mathbf{XB}$, onde **B** é a matriz de coeficientes da regressão:

$$\mathbf{B} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

2. Use a matriz $\hat{\mathbf{Y}}$ em uma análise de componentes principais de sua matriz de variância e covariância amostral padronizada, produzindo uma matriz **A** de autovetores. Os autovetores têm a mesma interpretação da PCA. Se você está fazendo uma RDA de uma matriz de características ambientais × espécies, os elementos de **A** são chamados de **escores das espécies.**
3. Gere dois conjuntos de escores a partir desta PCA.
   a. O primeiro conjunto de escores, **F**, é calculado multiplicando-se a matriz de autovetores **A** pela de variáveis respostas **Y**. O resultado, **F = YA**, é uma matriz em que as colunas são chamadas de **escores dos locais.** Os escores dos locais F são uma ordenação relativa a **Y**. Se você está usando uma matriz de características ambientais × espécies, os escores dos locais são fundamentados nas distribuições originais *observadas* das espécies.
   b. O segundo conjunto de escores, **Z**, é calculado multiplicando a matriz de autovetores **A** pela de valores ajustados $\hat{\mathbf{Y}}$. O resultado, **Z = YA**, é uma matriz na qual as colunas são chamadas **escores ajustados dos locais.** Como $\hat{\mathbf{Y}}$ também é igual à matriz **XB**, a matriz dos escores ajustados locais também pode ser escrita como **Z = XBA.** Os escores ajustados dos locais Z são uma ordenação relativa a **X.** Se você está usando uma matriz de características ambientais × espécies, os escores ajustados dos locais baseiam-se na distribuição das espécies que são *preditas* pelo ambiente.
4. Em seguida, queremos saber como as variáveis preditoras em **X** contribuem para cada um dos eixos de ordenação. A maneira mais simples de medir a contribuição das variáveis preditoras é observar que a matriz **Z** é composta de duas partes: **X**, os valores preditos, e **BA**, a matriz produto dos coeficientes de regressão e dos autovetores da matriz $\hat{\mathbf{Y}}$. A matriz **C = BA** nos diz a contribuição das variáveis **X** para a de escores ajustados dos locais **Z.** Essa decomposição equivale a dizer que os valores em cada uma das colunas de **C** são iguais aos coeficientes de regressão padronizados da matriz **X.** Um método alternativo é determinar a correlação entre as variáveis ambientais **X** (preditoras) e os escores dos locais **F**, que é o produto das variáveis respostas e seus autovetores.
5. Por fim, construímos um *bi-plot*, como fizemos na análise de correspondência, no qual plotamos os eixos principais dos escores ou de **F** (os observados) ou de **Z** (os ajustados) para as variáveis preditoras. Neste gráfico, podemos posicionar a localização das variáveis respostas. Desta forma, podemos visualizar a relação multivariada das variáveis respostas com as preditoras.

UM BREVE EXEMPLO Para ilustrar a RDA, retornamos aos dados de caramujos (Tabela 12.11). Para cada um dos locais também temos medidas de quatro variáveis ambientais: precipitação anual (mm), número de meses secos por ano, temperatura mensal média, e altura média do dossel do mangue no qual os caramujos forrageavam (Tabela 12.14). Os resultados dos quatro passos da RDA estão apresentados na Tabela 12.15. Como as variáveis respostas são medidas da forma da concha, a matriz **A** é composta de escores da forma da concha. As matrizes F e **Z** são escores dos locais e escores ajustados dos locais, respectivamente.

A Figura 12.15 ilustra como os locais covariam com as variáveis ambientais (símbolos verdes e setas direcionais) e como as formas das conchas covariam com os locais e suas variáveis ambientais (símbolos pretos e setas direcionais). Por exemplo, alguns locais no Brasil, na Nicarágua e na África são caracterizados por maior altura do dossel e maior precipitação, enquanto na Flórida e Caribe são caracterizados por maior temperatura e meses secos mais longos. As conchas do Caribe e da Flórida têm maior proporcionalidade e maior altura do espiral, enquanto as da África, do Brasil e da Nicarágua são mais circulares.

TESTE DA SIGNIFICÂNCIA DOS RESULTADOS Os resultados da RDA (e da CCA) podem ser testados em termos de significância estatística usando métodos de simulação de Monte Carlo (*ver* Capítulo 5). A hipótese nula é que não há nenhuma relação entre as variáveis preditoras e respostas. As randomizações dos

TABELA 12.14 Variáveis ambientais usadas na análise de redundância (RDA)

País	Precipitação anual (mm)	Nº de meses secos	Temp. média mensal (°C)	Altura média do dossel da floresta (m)
Angola	363	9	26,4	30
Bahamas	1.181	2	25,1	3
Belize	1.500	2	29,5	8
Brasil	2.150	4	26,4	30
Flórida	1.004	1	25,3	10
Haiti	1.242	6	27,5	10
Libéria	3.874	3	27,0	30
Nicarágua	3.293	0	26,0	15
Serra Leoa	4.349	4	26,6	35

Dados de Merkt & Ellison (1998).

TABELA 12.15 Análise de redundância da morfologia de pequenas conchas e dados ambientais

A. Calcule a matriz de valores ajustados $\hat{Y} = X(X^T X)^{-1} X^T Y$.

$$\hat{Y} = \begin{bmatrix} -0,05 & 0,01 & -0,04 \\ 0,09 & 0,00 & 0,09 \\ 0,00 & 0,01 & 0,08 \\ -0,05 & -0,01 & -0,08 \\ 0,05 & 0,00 & 0,03 \\ 0,04 & 0,02 & 0,09 \\ -0,05 & -0,01 & -0,07 \\ 0,03 & -0,01 & 0,00 \\ -0,06 & -0,01 & -0,10 \end{bmatrix}$$

Essa matriz é o resultado da regressão de cada uma das três variáveis de forma contra as quatro variáveis ambientais. Cada linha de $\hat{Y}$ é uma observação (de um local; *ver* Tabela 12.11), e cada coluna é uma das variáveis (proporcionalidade, circularidade e altura do espiral, respectivamente).

B. Use $\hat{Y}$ em uma PCA para calcular os autovetores A.

$$A = \begin{bmatrix} 0,55 & -0,81 & -0,19 \\ 0,08 & 0,27 & -0,96 \\ 0,83 & 0,52 & 0,21 \end{bmatrix}$$

Esses são os escores da forma da concha, e poderiam ser usados para calcular os valores dos componentes principais (Z_i) para a forma da concha. As colunas representam os três primeiros autovetores, as linhas representam as três variáveis morfológicas. O componente 1 explica 94% da variância na forma da concha. A circularidade da concha, que é grosseiramente constante entre os locais (*ver* Tabela 12.11), tem pouco peso ($a_{21} = 0,08$) no Componente Principal 1.

C. Calcule a matriz dos escores dos locais $F = YA$ e a matriz dos escores ajustados dos locais $Z = \hat{Y}A$.

$$F = \begin{bmatrix} 2,21 & -0,03 & -0,63 \\ 2,44 & -0,06 & -0,62 \\ 2,38 & 0,00 & -0,61 \\ 2,26 & -0,08 & -0,62 \\ 2,40 & -0,02 & -0,59 \\ 2,45 & -0,03 & -0,61 \\ 2,21 & -0,03 & -0,62 \\ 2,28 & -0,13 & -0,63 \\ 2,23 & -0,01 & -0,59 \end{bmatrix} \qquad Z = \begin{bmatrix} -0,06 & 0,02 & -0,01 \\ 0,13 & -0,03 & 0,00 \\ 0,07 & 0,04 & 0,01 \\ -0,09 & 0,00 & 0,00 \\ 0,05 & -0,02 & 0,00 \\ 0,10 & 0,02 & 0,00 \\ -0,09 & 0,00 & 0,00 \\ 0,01 & -0,03 & 0,00 \\ -0,12 & 0,00 & 0,00 \end{bmatrix}$$

A matriz F contém os escores dos componentes principais para os valores de forma da concha obtidos multiplicando os valores originais (Y) pelos autovetores da matriz A. A matriz Z contém os escores dos componentes principais para os valores de forma da concha obtidos multiplicando os valores ajustados ($\hat{Y}$), que são derivados da matriz de hábitat X (Tabela 12.14), pelos autovetores em A.

(*Continua*)

TABELA 12.15 Análise de redundância da morfologia de pequenas conchas e dados ambientais (*Continuação*)

D. Calcule a matriz C = BA para obter os coeficientes da regressão contra X.

$$\mathbf{C} = \begin{bmatrix} 0{,}01 & 0{,}00 & 0{,}00 \\ 0{,}03 & 0{,}01 & 0{,}00 \\ 0{,}00 & 0{,}02 & 0{,}00 \\ -0{,}11 & 0{,}00 & 0{,}00 \end{bmatrix}$$

As linhas são as quatro variáveis ambientais e as colunas, os três primeiros componentes principais. Os coeficientes 0,00 são valores arredondados de números menores. Alternativamente, podemos calcular as correlações das variáveis ambientais X com os escores dos locais F. As entradas das células são os coeficientes de correlação entre X e as colunas de F:

	Componente 1	Componente 2	Componente 3
Precipitação	–0,55	–0,19	0,11
Meses secos	–0,28	0,34	–0,18
Temperatura média	0,02	0,49	0,11
Altura do dossel	–0,91	0,04	–0,04

Neste exemplo, regredimos as características da forma da concha de caramujos (Tabela 12.11) contra dados ambientais (Tabela 12.13). Antes da análise, os dados ambientais foram padronizados usando a Equação 12.17. (Dados de Merkt & Ellison, 1998.)

dados brutos – intercambiando as linhas da matriz Y – são seguidas de repetição da RDA. Então, comparamos a estatística-F calculada dos dados originais com a distribuição das estatísticas-F dos dados randomizados. A estatística F para esse teste é:

$$F = \frac{\left(\frac{\sum_{j=1}^{n} \lambda_j}{m} \right)}{\left(\frac{SQR}{(p-m-1)} \right)} \qquad (12.32)$$

Nesta equação, $\sum \lambda_j$ é a soma dos autovalores, SQR é a soma dos quadrados dos resíduos (calculada como a soma dos autovalores da PCA usando $(\mathbf{Y} - \hat{\mathbf{Y}})$ no lugar de $\hat{\mathbf{Y}}$ (Tabela 12.15B), m é o número de variáveis preditoras e p é o número de observações. Embora o tamanho da amostra seja muito pequeno para uma análise significativa, ainda podemos usar a Equação 12.32 para ver se nossos resultados são estatisticamente significativos. O valor de F para nossos dados originais = 3,27. Um histograma dos valores de F de 1.000 randomizações dos nossos dados é mostrado na Figura 12.16. A comparação do valor de F observado com as

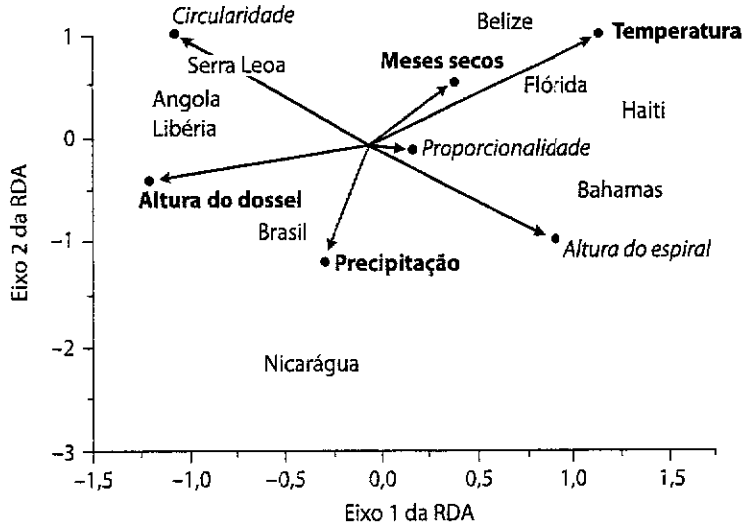


Figura 12.15 *Bi-plot* dos dois primeiros eixos de uma Análise de Redundância (RDA) da regressão dos dados de conchas de caramujos (Tabela 12.11) contra os dados ambientais (Tabela 12.14), sendo que todos eles foram medidos em 9 países (Merkt & Ellison, 1998). Os nomes dos países indicam a localização de cada local no espaço de ordenação (a matriz **F** da Tabela 12.15). Os símbolos pretos indicam a localização das variáveis morfológicas no espaço dos locais. Então, as conchas de Belize, Flórida, Haiti e Bahamas têm maior proporcionalidade e maior altura do espiral que as dos outros locais, enquanto as conchas da África, do Brasil e da Nicarágua são mais circulares. As setas pretas indicam a direção do aumento da variável morfológica. Similarmente, os símbolos e setas verdes indicam como os locais foram ordenados. Todos os valores foram padronizados para permitir a plotagem de todas as matrizes em escalas similares. Então, os comprimentos das setas não indicam a magnitude dos efeitos. Entretanto, indicam sua direção ou seu peso sobre os dois eixos.

randomizações dá um valor de P exato de 0,095, que é marginalmente maior do que o tradicional ponto de corte de $P = 0,05$. A conclusão limitada (com base em um tamanho amostral limitado) é que não existe relação significativa entre morfologia do caramujo e variação ambiental entre esses nove locais.[13] *Ver* Legendre & Legendre (1998) para detalhes adicionais e exemplos de testes de hipóteses na RDA e na CCA.

Uma extensão da RDA, aquela com base em distâncias (ou db-RDA) é um método alternativo para testar modelos multivariados complexos (Legendre & Anderson, 1999; McArdle & Anderson, 2001). Ao contrário da MANOVA clássica, a db-RDA não requer que os dados tenham uma distribuição normal multivariada, que as medidas de distância entre observações sejam euclidianas, ou que existam mais observações do que variáveis respostas medidas. A RDA com base em distâncias pode ser usada com qualquer medida de distância, métrica ou semimétrica, e permite a partição dos componentes de variação em modelos

[13] Em contraste, a análise do conjunto de dados completo com 1.042 observações de 19 países deu uma associação significativa entre a forma da concha do caramujo e as condições ambientais (Merkt & Ellison, 1998).

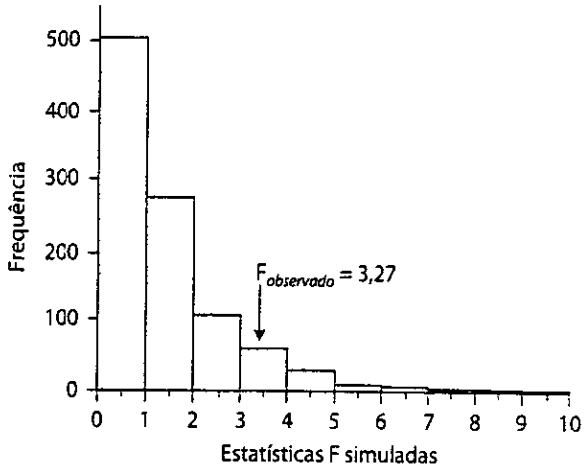


Figura 12.16 Distribuição de frequência das estatísticas F resultantes de 1.000 randomizações e análise de redundância (RDA) dos dados de caramujos (Tabela 12.1'), em que os rótulos dos países foram embaralhados de maneira aleatória entre as amostras. A estatística F para os dados observados = 3,27. Noventa e cinco randomizações tiveram uma estatística F maior, de forma que o valor de *P* para a RDA na Figura 12.15 = 0,095. Esse resultado é marginalmente não significativo usando um nível de α igual a 0,05, refletindo a limitação de usar a RDA em um conjunto de dados tão pequeno.

de MANOVA complexos. Como a RDA, ela usa testes de randomização para determinar a significância dos resultados. A RDA com base em distâncias tem sido implementada nas novas versões de programas para análise multivariada (*ver* Rejmánek & Klinger, 2003, para uma revisão recente desses programas).

RESUMO

Os dados multivariados contêm múltiplas variáveis respostas não independentes medidas para cada réplica. A análise de variância multivariada (MANOVA) e a de redundância (RDA) são as análogas multivariadas dos métodos de ANOVA e regressão usados em análises univariadas. Os testes estatísticos com base na MANOVA e na RDA assumem amostragem aleatória e independente, assim como na análise univariada, mas apenas a MANOVA requer que os dados estejam de acordo com uma distribuição normal multivariada.

A "ordenação" é um conjunto de métodos multivariados designados para ordenar as amostras ou observações de dados multivariados, e para reduzir dados multivariados para um número menor de variáveis para análises adicionais. A distância euclidiana é a medida de distância mais natural entre amostras no espaço multivariado, mas essa medida pode dar resultados contraintuitivos em conjuntos de dados com muitos zeros, como os de abundância ou de presença e ausência de espécies. Outras medidas de distância, incluindo as de Manhattan e de Jaccard, têm propriedades algébricas melhores, embora índices de similaridade biogeográfica com base nessas medidas sejam muito sensíveis a efeitos do tamanho amostral.

A análise de componentes principais (PCA) é um dos métodos mais simples para ordenação de dados multivariados, mas não é útil, a menos que existam correlações subjacentes nos dados originais. A PCA extrai novas variáveis ortogonais que capturam a variação importante dos dados e preserva as distâncias das amostras originais no espaço multivariado. A PCA é um caso especial de análise de coordenadas principais (PCoA). Enquanto a PCA é fundamentada em distâncias euclidianas, a PCoA pode ser usada com qualquer medida de distância. A análise fatorial é um tipo de PCA inversa, na qual as variáveis medidas são decompostas em combinações lineares de fatores subjacentes. Como os resultados de uma análise fatorial são sensíveis aos métodos de rotação e como o mesmo conjunto de dados pode gerar mais de um conjunto de fatores, a análise fatorial apresenta menor utilidade que a PCA.

A análise de correspondência (CA) é uma ferramenta de ordenação que revela as associações entre as assembleias de espécies e as características dos locais. Ela ordena simultaneamente as espécies e os locais, e pode revelar agrupamentos ao longo de ambos os eixos. Entretanto, muitos problemas têm sido identificados na correlata análise de correspondência destendenciada (DCA), que foi desenvolvida para lidar com o efeito do arco comum na CA (e outras técnicas de ordenação). Uma alternativa à DCA é fazer uma CA usando outras medidas de distância. O escalonamento multidimensional não métrico (NMDS) preserva a ordem de ranque das distâncias, mas não as próprias, entre as variáveis no espaço multivariado. Os métodos de ordenação são úteis para redução de dados e para revelar padrões nos dados. Os escores da ordenação podem ser usados em métodos padrão para testes de hipóteses, desde que os seus pressupostos sejam atendidos.

Os métodos de classificação são usados para agrupar objetos em classes que possam ser identificadas e interpretadas. A análise de agrupamentos usa dados multivariados para gerar grupos de amostras. Os algoritmos de agrupamento hierárquico ou são aglomerativos (observações separadas são sequencialmente agrupadas em grupos) ou são divisivos (todo o conjunto de dados é sequencialmente dividido em grupos). Os métodos não hierárquicos, como o agrupamento por *K*-médias, requerem que o usuário especifique o número de grupos que serão criados, embora regras estatísticas de parada possam ser usadas para decidir o número correto de grupos. A análise discriminante é um tipo de análise de agrupamento ao contrário: os grupos são definidos *a priori*, e a análise produz escores das amostras que dão a melhor classificação. A randomização por jackknife e o teste de conjuntos de dados independentes podem ser usados para avaliar a confiabilidade da função discriminante. Os métodos de classificação assumem, *a priori*, que os grupos ou agrupamentos são relevantes biologicamente, enquanto o teste de hipóteses clássico começaria com uma hipótese nula de variação aleatória e nenhum grupo ou agrupamento subjacente.

Apêndice

Álgebra de Matrizes para Ecólogos

Muitas técnicas estatísticas podem ser melhor ilustradas usando álgebra de matrizes. Em particular, as técnicas multivariadas descritas no Capítulo 12 recaem sobre álgebra de matrizes, e os métodos de regressão (Capítulo 9) e ANOVA (Capítulo 10) frequentemente são apresentados em notação de matrizes. Se você entender o básico da álgebra de matrizes, poderá fazer programas simples para implementar novos métodos estatísticos que são apresentados em artigos científicos. Neste apêndice, explicamos, com brevidade, os fundamentos da álgebra de matrizes. Se você deseja aprender mais sobre álgebra de matrizes em estatística, Harville (1997) é uma referência excelente e fácil de ler.

O QUE É UMA MATRIZ?

A definição básica de matriz é uma tabela de números em que as linhas são as réplicas e as colunas são as variáveis medidas. Mais formalmente, uma **matriz** é um arranjo de regular de números que possui m linhas e n colunas. As matrizes são escritas de forma expandida com coeficientes para cada registro. Na forma compacta, uma matriz é simbolizada por uma letra maiúscula em negrito:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Cada **elemento** da matriz é indexado com dois subscritos: o primeiro indica a linha e o segundo, a coluna. Dessa forma, a_{34} é o elemento da matriz A que está na terceira linha e quarta coluna.

A **dimensão** de uma matriz é um par de números que especifica o número de linhas e de colunas de uma matriz:

$$\text{Dim}(\mathbf{A}) = (m, n)$$

As matrizes podem ter qualquer número de linhas ou colunas, mas três tamanhos são importantes particularmente. A **matriz quadrada** é aquela na qual o número de linhas é igual ao de colunas ($m = n$):

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

A soma dos elementos de uma matriz quadrada que estão na diagonal principal, $a_{11} + a_{22} + \ldots + a_{nn}$, é chamada de **traço** da matriz.

Um **vetor coluna** é uma matriz com m linhas e uma coluna:

$$\mathbf{A} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix}$$

Um **vetor linha** é uma matriz com uma linha e n colunas:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \end{bmatrix}$$

Na álgebra de matrizes, um único número k é chamado de escalar, para distingui-lo de uma matriz com uma linha e uma coluna:

$$\mathbf{A} = [a_{11}]$$

OPERAÇÕES MATEMÁTICAS BÁSICAS COM MATRIZES

Adição e subtração

A soma e a subtração de matrizes são feitas elemento por elemento. Para duas matrizes A e B com elementos $\{a_{ij}\}$ e $\{b_{ij}\}$, respectivamente, sua soma A + B é uma nova matriz C em que o elemento $i\,j$ é igual a $a_{ij} + b_{ij}$:

$$\begin{bmatrix} 3 & 4 \\ -1 & 5 \\ 0 & 2 \end{bmatrix} + \begin{bmatrix} 4 & 8 \\ 1 & -3 \\ -2 & 6 \end{bmatrix} = \begin{bmatrix} 3+4 & 4+8 \\ -1+1 & 5+(-3) \\ 0+(-2) & 2+6 \end{bmatrix} = \begin{bmatrix} 7 & 12 \\ 0 & 2 \\ -2 & 8 \end{bmatrix} \qquad \text{(A.1)}$$

De forma similar, a diferença entre duas matrizes A – B é uma nova matriz C em que o elemento $i\,j$ é igual a $a_{ij} - b_{ij}$. Como a adição e a subtração de matrizes são feitas elemento por elemento, apenas matrizes com a mesma dimensão (i. e., matrizes com o mesmo número de linhas e de colunas) podem ser somadas ou subtraídas uma da outra.

A adição de matrizes é **comutativa**, o que significa que, para duas matrizes A e B de mesma dimensões, A + B = B + A. A adição de matrizes também é **associa-**

tiva: para três matrizes A, B e C com dimensões iguais, A + (B + C) = (A + B) + C. Você já deve ter visto essas propriedades comutativas e associativas antes, pois elas também se aplicam à aritmética simples.

Multiplicação

Dois tipos de cálculo são possíveis com matrizes. Quando multiplicamos uma matriz A de qualquer dimensão por um único número, ou escalar k, o produto $k\mathbf{A}$ é uma nova matriz com a mesma dimensão de A em que os elementos $i\,j$ são iguais a ka_{ij}:

$$3\begin{bmatrix}1 & 3 & 2 & -5\\ 2 & 0 & 0 & 3\end{bmatrix}=\begin{bmatrix}3\times1 & 3\times3 & 3\times2 & 3\times(-5)\\ 3\times2 & 3\times0 & 3\times0 & 3\times3\end{bmatrix}=\begin{bmatrix}3 & 9 & 6 & -15\\ 6 & 0 & 0 & 9\end{bmatrix} \quad \textbf{(A.2)}$$

Como na adição e subtração de matrizes, a multiplicação por uma escalar é comutativa ($k\mathbf{A} = \mathbf{A}k$) e associativa ($c(k\mathbf{A}) = (ck)\mathbf{A}$). A multiplicação escalar também é **distributiva** em respeito à adição: $k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B}$. Como na aritmética simples, $1 \times \mathbf{A} = \mathbf{A}$.

Também é possível multiplicar duas matrizes A e B, com dimensões (m, n) e (p, q), respectivamente, mas apenas se os números de colunas da primeira matriz, A, for igual ao número de linhas da segunda matriz, B (i. e., se $n = p$). O produto AB de duas matrizes é uma nova matriz, cujo elemento $i\,j$ é igual a

$$\sum_{k=1}^{n} a_{ik}b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{in}b_{nj}$$

Por exemplo:

$$\begin{bmatrix}1 & 3 & 3\\ 2 & 0 & -2\end{bmatrix}\begin{bmatrix}1 & -2 & 2\\ 3 & 2 & 2\\ -1 & 3 & 3\end{bmatrix}$$

$$=\begin{bmatrix}1(1)+3(3)+3(-1) & 1(-2)+3(2)+3(3) & 1(2)+3(2)+3(3)\\ 2(1)+0(3)+(-2)(-1) & 2(-2)+0(2)+(-2)(3) & 2(2)+0(2)+(-2)(3)\end{bmatrix}$$

$$=\begin{bmatrix}7 & 13 & 17\\ 4 & -10 & -2\end{bmatrix} \quad \textbf{(A.3)}$$

A dimensão de AB = (m, q), onde m é o número de linhas da matriz A e q é o de colunas da matriz B.

Para três matrizes A, B e C, a multiplicação de matrizes é associativa [A(BC) = (AB)C], e é distributiva em relação à adição [A(B + C) = AB + AC e

$(A + B)C = AC + BC$]. A multiplicação de matrizes em geral não é comutativa, contudo: **AB** muitas vezes não é igual a **BA**. De fato, você nem sempre pode multiplicar um par de matrizes em ambas direções, a menos que $n = p$ e $m = q$. Se ambas, **A** e **B**, são matrizes quadradas com dimensões iguais ($m = n = p = q$), AB e BA são definidas com (Dim(**AB**) = (n,n)), mas nem sempre é o caso em que **AB** = **BA**. Por exemplo:

$$\begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}\begin{bmatrix} 0 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 8 & 4 \end{bmatrix} \neq \begin{bmatrix} 0 & 8 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \quad \text{(A.4)}$$

Chamamos **AB** de **pré-produto** ou a **pré-multiplicação** de **B** por **A**. De forma equivalente, dizemos que **AB** é o **pós-produto** ou **pós-multiplicação** de **A** por **B**.

Note que há uma diferença entre multiplicar uma matriz por uma escalar $k\mathbf{A}$ e uma matriz por uma matriz com uma linha e uma coluna $[k]\mathbf{A}$. A primeira operação é possível para uma matriz com qualquer número de linhas e colunas (Equação A.2), mas a segunda é possível apenas para uma matriz com uma linha.

Se multiplicarmos um vetor linha [dimensões = $(1,n)$] por um vetor coluna [dimensões = $(p,1)$] o resultado será uma matriz 1×1. Esse resultado em geral é chamado de **produto escalar** (pois há apenas um elemento no produto desses dois vetores) e muitos autores o tratam como uma escalar (p. ex., Legendre & Legendre, 1998), mas a maioria dos programas estatísticos irá tratar esse resultado como uma matriz.

Dois vetores cujo produto é igual a 0 são chamados de **ortogonais**. Por exemplo, um par de contrastes *a priori* ortogonais da ANOVA (*ver* Capítulo 10) são dois vetores ou números, tal que a soma de seus produtos é igual a 0. Em termos de álgebra de matrizes, o par de contrastes representa um vetor linha e um vetor coluna cujo produto das matrizes é 0, e portanto eles são ortogonais.

Transposição

Intercambiar as linhas e colunas da matriz produz sua **transposta**. Para uma matriz **A** com dimensões (m,n), a transposta de **A**, escrita como **A**′ ou $\mathbf{A}^T$, tem dimensões (n,m) cujo elemento $i\,j$ é igual ao elemento $j\,i$ da matriz **A**:

$$\begin{bmatrix} 3 & 0 \\ -1 & -2 \\ 2 & -1 \end{bmatrix}^T = \begin{bmatrix} 3 & -1 & 2 \\ 0 & -2 & -1 \end{bmatrix} \quad \text{(A.5)}$$

A transposição de uma matriz transposta recupera a matriz original: $(\mathbf{A}^T)^T = \mathbf{A}$.

Chamamos uma matriz de **simétrica** se ela não muda após sua transposição: $\mathbf{A}^T = \mathbf{A}$. Somente matrizes quadradas podem ser simétricas. Os elementos de uma matriz quadrada simétrica são os mesmos nos dois lados da diagonal:

$$\begin{bmatrix} 1 & 2 & 2 & -1 \\ 2 & 4 & 3 & 2 \\ 2 & 3 & -4 & 0 \\ -1 & 2 & 0 & 8 \end{bmatrix}$$

Outro exemplo de matriz quadrada simétrica é a de variância e covariância (*ver* Capítulo 9, Nota de rodapé 5). Para um conjunto de n variáveis, a matriz de variância covariância $n \times n$ contém a covariância σ_{ij} entre as variáveis i e j, com a variância da variável i, σ_{ii}, como elementos da diagonal. Como a covariância de $\sigma_{ij} = \sigma_{ji}$, os elementos dos dois lados da diagonal são idênticos e a matriz é simétrica.

Existem duas matrizes simétricas que são especialmente úteis. A **matriz diagonal D** é do tipo quadrada, na qual todos os elementos são iguais a 0, exceto aqueles na diagonal principal.

$$\mathbf{D} = \begin{bmatrix} a_{11} & 0 & 0 & 0 & 0 \\ 0 & a_{22} & 0 & 0 & 0 \\ 0 & 0 & a_{33} & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & a_{nn} \end{bmatrix} \tag{A.6}$$

A matriz diagonal, cujos elementos são iguais a 1, é chamada de **matriz identidade**, e geralmente escrita como **I**:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{A.7}$$

Uma matriz é chamada de **triangular** se todos os elementos em um dos lados da diagonal são iguais a 0. Se todos os elementos acima da diagonal são iguais a zero, ela se chama **matriz triangular inferior**:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 13 & 2 & 0 & 0 \\ 2 & -1 & 2 & 0 \\ 4 & 2 & -2 & -1 \end{bmatrix} \tag{A.8}$$

Em contraste, se todos os elementos abaixo da diagonal forem iguais a zero, ela é chamada de **matriz diagonal superior**:

$$A = \begin{bmatrix} 1 & 10 & 3 & -2 \\ 0 & 2 & 3 & 5 \\ 0 & 0 & 2 & -3 \\ 0 & 0 & 0 & -1 \end{bmatrix} \tag{A.9}$$

Note que a transposição de uma matriz triangular inferior resulta em uma triangular superior, e vice-versa.

INVERSÃO Para uma simples variável ou escalar k, sabemos que seu inverso k^{-1} é definido tal que $k \times k^{-1} = 1$. Por analogia, para uma matriz A, definimos seu inverso por A^{-1} como sendo a matriz que satisfaz a equação $A \times A^{-1} = I$, a matriz identidade. As matrizes inversas são definidas somente para quadradas, mas nem todas as matrizes quadradas possuem inversos. Se A^{-1} existe, então $AA^{-1} = A^{-1}A = I$. Um exemplo de **matriz inversa** é:

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} -2 & 1{,}5 \\ 1 & -0{,}5 \end{bmatrix} \tag{A.10}$$

Você pode verificar que essas duas matrizes são invertidas pela multiplicação entre elas. Veja que seu produto é a matriz identidade I. Para os casos especiais, nos quais o inverso de uma matriz é igual a sua transposta, $A^{-1} = A^T$, a matriz A é chamada de **ortogonal**. O cálculo de uma matriz inversa é melhor realizado com um programa de matemática ou de estatística.

FORMAS QUADRÁTICAS Se combinarmos a transposição com a multiplicação de matrizes de uma forma particular, obtemos uma expressão chamada de **forma quadrática**. Dada uma matriz quadrada A de dimensões (n,n) e um vetor coluna de dimensões (n,1), a expressão:

$$Q = x^T A x \tag{A.11}$$

é chamada de forma quadrática. Assim, Q é uma matriz 1×1 com seu elemento igual a:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j \tag{A.12}$$

Q geralmente é considerada uma escalar, embora alguns programas matemáticos e estatísticos possam tratá-la como uma matriz 1×1.

RESOLVENDO SISTEMAS DE EQUAÇÕES LINEARES Uma das mais importantes aplicações da álgebra de matrizes é para resolver sistemas de equações lineares. Em particular, a estimativa dos parâmetros β_0 e β_1 de um modelo básico de regressão linear (*ver* Capítulo 9) equivale a resolver um sistema de equações lineares:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 \\ Y_3 &= \beta_0 + \beta_1 X_3 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n \end{aligned} \qquad \textbf{(A.13)}$$

Se **Y** é o vetor coluna:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

que representa cada uma das n respostas; **X** é uma matriz $n \times 2$:

$$\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & \vdots \\ 1 & X_n \end{bmatrix}$$

que representa cada um dos preditores; e **b** é o vetor coluna:

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

de parâmetros para o modelo de regressão linear, então o sistema de equações lineares (Equação A.13) pode ser representado como $\mathbf{Y} = \mathbf{Xb}$. Os valores de β_0 e β_1 que fornecem o melhor ajuste para o modelo de regressão (o modelo de mínimos quadrados descrito no Capítulo 9) pode ser calculado como:

$$\mathbf{b} = [\mathbf{X}^T\mathbf{X}]^{-1}[\mathbf{X}^T\mathbf{Y}] \qquad \textbf{(A.14)}$$

Esta equação funciona igualmente bem tanto para a regressão simples como para a regressão múltipla.

AUTOVALORES E AUTOVETORES Um conjunto especial de equações lineares:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= \lambda x_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= \lambda x_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= \lambda x_n \end{aligned} \tag{A.15}$$

pode ser escrito em forma de matrizes como:

$$\mathbf{Ax} = \lambda\mathbf{x} \tag{A.16}$$

onde **A** é uma matriz quadrada com elementos $a_{i,j}$, **x** é um vetor coluna $n \times 1$, e λ é uma escalar. Usando as regras de álgebra de matrizes até então, a Equação A.16 pode ser re-expressa como:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \tag{A.17}$$

onde **0** é um vetor coluna cujos elementos são todos 0. Se a Equação A.17 (ou A.16 ou a A.15) pode ser resolvida (o que nem sempre é o caso), aplicar-se-á somente para certos valores de λ. Tais valores são chamados de **autovalores** da matriz **A**. Como existem n equações na Equação A.15, pode haver até n autovalores, que podem ser números complexos (constantemente) ou números reais (de vez em quando). Para cada um desses autovalores λ_i, há um **autovetor** correspondente, que é o vetor coluna **x**, que satisfaz a Equação A.16 ou a A.17. Por fim, a soma dos autovalores da matriz **A** é igual ao seu *traço*.

DETERMINANTES A equação para calcular os autovalores (Equação A.17) tem uma solução trivial quando $\mathbf{x} = \mathbf{0}$ (um vetor coluna com todos os elementos iguais a zero). Em adição, a Equação A.17 tem outra solução:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0 \tag{A.18}$$

As duas linhas verticais ($|\ |$) na Equação A.18 se referem ao determinante da diferença entre as matrizes **A** e λ**I**. O determinante de uma matriz quadrada **A** é uma escalar, que é calculada a partir da soma de todos os produtos com sinais possíveis (positivos e negativos), contendo um elemento da linha e um elemento da coluna de **A**. O sinal de cada produto é determinado por uma regra que primeiro irá parecer arbitrária.

A matriz quadrada **A** tem elementos a_{ij}. Para qualquer produto entre cada par de elementos a_{ij} e a_{km}, onde $i \neq k$ e $j \neq m$, definimos o produto $a_{ij}a_{km}$ como um *par negativo* se um dos elementos estiver posicionado abaixo e à direita do outro. Este será o caso se $k > i$ e $m < j$. Caso contrário, o produto será um *par positivo*. A atribuição do sinal pode ser visualizada usando a tabela a seguir:

	$k > i$	$k < i$
$m > j$	+	−
$m < j$	−	+

Por exemplo, em uma matriz $n \times n$ onde $n > 4$, o par a_{34} e a_{22} é positivo, e o par a_{34} e a_{41} é negativo. A atribuição dos sinais positivos e negativos aos pares está fundamentada na posição relativa dos dois elementos dentro da matriz e não tem qualquer relação com o sinal real (positivo ou negativo) dos elementos ou de seu produto.

Existem n elementos da matriz quadrada A, que dois deles caíam ou na mesma linha ou na mesma coluna. Um exemplo deles são os $i_1j_1, \ldots, i_nj_n$ elementos de A, onde $i_1, \ldots, i_n$ e $j_1, \ldots, j_n$ são permutações dos primeiros n números inteiros $(1, \ldots, n)$. Um total de $\binom{n}{2}$ pares podem ser formados a partir desses n elementos.

Usamos o símbolo $\sigma_n(i_1,j_1;\ldots;i_n,j_n)$ para representar o número desses pares negativos.

O determinante da matriz quadrada A, escrito como $|A|$, é definido como:

$$|A| = \sum_{i=1}^{n} (-1)^{\sigma_n(i_1,j_1;\ldots;i_n,j_n)} a_{1j_1} \ldots a_{nj_n} \qquad \text{(A.19)}$$

onde $j_1, \ldots, j_n$ é uma permutação dos n primeiros números inteiros e o somatório é sobre todas as permutações.

A receita para calcular o determinante é:

1. Compare todos os produtos possíveis de cada um dos n fatores que podem ser obtidos através da escolha de um, e somente um, elemento de cada linha e coluna da matriz quadrada A.
2. Para cada produto, conte o número de pares negativos entre os $\binom{n}{2}$ pares que podem ser encontrados para os n elementos desse produto. Se par, o sinal do produto será +. Se o número de pares negativos for um número impar, o sinal do produto será −.
3. Some os produtos com sinal.

Para uma matriz 2×2, o determinante é:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21} \qquad \text{(A.20)}$$

Para uma matriz 3 × 3, o determinante é:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \quad \text{(A.21)}$$

Como os determinantes podem ser calculados usando tanto as linhas como as colunas de uma matriz, o da matriz A é igual ao de sua transposta: $|\mathbf{A}| = |\mathbf{A}^T|$. Uma propriedade útil de uma matriz triangular é que seu determinante é igual ao produto dos elementos da diagonal principal (já que todos os outros produtos na Equação A.19 = 0).

Em adição à explicação do mecanismo para identificar autovalores (Equação A.18), os determinantes têm outras propriedades úteis:

1. Os determinantes transformam o resultado da multiplicação de matrizes em uma escalar: $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$.
2. A matriz quadrada cujo determinante é igual a zero não tem inverso, e é chamada de **matriz singular**.
3. O determinante da matriz é igual ao produto de seus autovalores:

$$|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i$$

4. Para qualquer escalar k e qualquer matriz $n \times n$, A, $|k\mathbf{A}| = k^n|\mathbf{A}|$.

Como visto no item 2, nem todas as matrizes possuem inverso. Em particular, o inverso de uma matriz $\mathbf{A}^{-1}$, apenas existe se o determinante de A, $|\mathbf{A}|$, não for igual a 0.

DECOMPOSIÇÃO DE VALOR SINGULAR Uma propriedade útil de que qualquer matriz A, $m \times n$, pode ser re-escrita, ou **decomposta**, usando a fórmula a seguir:

$$\mathbf{A} = \mathbf{VWU}^T \quad \text{(A.22)}$$

onde V tem as mesmas dimensões de A ($m \times n$); U é uma matriz $n \times n$; V e U são **matrizes com colunas ortogonais**: as matrizes ortogonais, cujas colunas foram normalizadas de forma que a distância euclidiana de cada coluna é

$$\sqrt{\sum_{k=1}^{n} (y_{ik} - y_{jk})^2} = 1$$

e W é uma matriz diagonal $n \times n$ cujos valores da diagonal w_{ii} são chamados de **valores singulares** de A. A Equação A.22, conhecida como **decomposição de valores singulares**, pode ser usada para determinar se uma matriz quadrada A é uma

matriz singular. Por exemplo, para resolver o sistema de equações lineares, (A.14) requer a avaliação da expressão $\mathbf{Y} = \mathbf{Xb}$, onde $\mathbf{X}$ é uma matriz quadrada. Como mostramos na equação A.14, a solução de $\mathbf{b}$ requer tirar o inverso de $\mathbf{X}^T\mathbf{X}$. Para simplificar o restante da discussão, iremos escrever $\mathbf{X}^*$ para representar $\mathbf{X}^T\mathbf{X}$.

Usando a Equação A.22, podemos representar $\mathbf{X}^*$ como $\mathbf{VWU}^T$. Como o inverso do pré-produto de duas matrizes $[\mathbf{AB}]^{-1}$ é igual ao pós-produto de seus inversos $\mathbf{A}^{-1}\mathbf{B}^{-1}$, podemos calcular o inverso de $\mathbf{X}^*$ como:

$$\mathbf{X}^{*-1} = [\mathbf{VWU}^T]^{-1} = [\mathbf{U}^T]^{-1}\mathbf{W}^{-1}\mathbf{V}^{-1} \quad \textbf{(A.23)}$$

Como $\mathbf{U}$ e $\mathbf{V}$ são ortogonais, seus inversos são iguais às suas transpostas (que são bem mais fáceis de calcular), e:

$$\mathbf{X}^{*-1} = \mathbf{UW}^{-1}\mathbf{V}^T \quad \textbf{(A.24)}$$

Além disso, o inverso da matriz diagonal é apenas uma nova matriz diagonal, cujos elementos são os inversos do original. Se qualquer elemento da diagonal de $\mathbf{W} = 0$, então seus inversos são infinitos ($1/0 = \infty$) e o inverso de $\mathbf{X}^*$ não pode ser definido. Assim, podemos determinar com mais facilidade quando $\mathbf{X}^*$ é singular. Ainda é possível achar solução para a equação $\mathbf{Y} = \mathbf{Xb}$ quando $\mathbf{X}^T\mathbf{X}$ for singular, usando métodos computacionalmente intensivos (Press et al., 1986).

Glossário

Os números entre colchetes indicam o(s) capítulo(s) nos quais acontece a discussão mais completa desses termos.

abscissa (*abscissa*) A linha horizontal, ou eixo-*x*, em um gráfico. O oposto de **ordenada.** [7]

acurácia (*accuracy*) O grau em que a medida de um objeto se aproxima do seu valor real. *Comparar com* **precisão.** [2]

agrupamento aglomerativo (*agglomerative clustering*) O processo de agrupar objetos pegando grupos menores e os agrupando em grupos maiores, com base em sua similaridade. *Ver também* **análise de agrupamento**. *Comparar com* **agrupamento divisivo**. [12]

agrupamento divisivo (*divisive clustering*) O processo de agrupar objetos a partir da divisão de um grande grupo, com base na dissimilaridade, em outros sucessivamente menores. *Ver também* **análise de agrupamentos**. *Comparar com* **agrupamento aglomerativo**. [12]

agrupamento hierárquico (*hierarchical clustering*) Um método para agrupar objetos que se baseia na imposição de um sistema de referência externo. Em agrupamentos hierárquicos, se *a* e *b* estão em um grupo e as observações *c* e *d* estão em outro um grupo maior que inclui *a* e *c*, também deve incluir *b* e *d*. A taxonomia linneana é um exemplo de um sistema de agrupamento hierárquico. [12]

agrupamento não hierárquico (*non-hierarchical clustering*) Um método para agrupar objetos que não é fundamentado em um sistema de referência externo. *Comparar com* **agrupamento hierárquico**. [12]

agrupamento por K-médias (*K-means clustering*) Um método não hierárquico de agrupar objetos que minimiza a soma dos quadrados das distâncias de cada objeto até o centroide de seu grupo. [12]

ajuste (*fit*) Quão consistentes as predições são em relação ao observado. Usado para descrever quão bem os dados são preditos por um dado modelo, ou quão bem os modelos são preditos pelos dados disponíveis. [9, 10, 11, 12]

alfa (*alpha*) A primeira letra (α) do alfabeto grego. Na estatística, ela é usada para denotar uma probabilidade aceitável de cometer um erro Tipo I. [4]

amostra (*sample*) Um subconjunto da população de interesse. Como é praticamente impossível observar ou manipular todos os indivíduos em uma população, as análises baseiam-se em amostras da população. A probabilidade e a estatística são utilizadas para extrapolar os resultados a partir de uma amostra pequena para uma população maior. [1]

análise de agrupamento (*cluster analysis*) Um método para agrupar objetos com base em distâncias multivariadas entre eles. [12]

análise de caminhos (*path analysis*) Um método para atribuir causas e efeitos múltiplos e sobrepostos em um tipo especial de modelo de regressão múltipla. [9]

análise de componentes principais (PCA – *principal component analysis*) Um método de ordenação que reduz dados multivariados a um número menor de variáveis pela criação de combinações lineares das variáveis originais. Foi originalmente desenvolvida para atribuir identidade racial a indivíduos com base em múltiplas medidas biométricas. [12]

análise de coordenadas principais (PCoA – *Principal coordinate analysis*) Um método de ordenação genérico que pode usar qualquer medida de distância. A análise de componentes principais e a análise fatorial (por exemplo) são casos especiais de uma análise de coordenadas principais. [12]

análise de correspondência (*correspondence analysis*) Um método de ordenação usado para relacionar a distribuição de espécies a variáveis ambientais. Também chamada de "análise de gradiente indireta". [12]

análise de correspondência canônica (CCA – *canonical correspondence analysis*) Um método de ordenação, tipicamente usado para relacionar a abundância ou as características de espécies com variáveis

ambientais ou do hábitat. Esse método ordena linearmente os preditores (variáveis ambientais ou do hábitat) e as respostas (dados das espécies) para produzir eixos unimodais, ou de corcova, relativos aos preditores. Ela também é um tipo de análise de regressão multivariada. *Comparar com* a **análise de redundância.** [12]

análise de correspondência destendenciada (*detrended correspondence analysis*) Uma variação da análise de correspondência usada para melhorar (remover) o efeito do arco. [12]

análise de covariância (ANCOVA – *analysis of covariance*) Um modelo estatístico híbrido, algo entre regressão e ANOVA, que inclui uma variável preditora contínua (a covariável) em conjunto a uma variável preditora categórica de um modelo de ANOVA. [10]

análise de dados gráfica exploratória (EDA – *graphical exploratory data analysis*) O uso de estruturas visuais, como gráficos, para detectar padrões, dados discrepantes e erros nos dados. [8]

análise de intervenção randomizada (*randomized intervention analysis*) Uma técnica de Monte Carlo para a análise de delineamentos ADCI. [7]

análise de Monte Carlo (*Monte Carlo analysis*) Umas das três grandes estruturas de análises estatísticas. Ela usa métodos de Monte Carlo para estimar valores de *P*. *Ver* métodos de Monte Carlo. *Comparar com* **inferência bayesiana** e **análises paramétricas.** [5]

análise de redundância (RDA – *redundance analysis*) Um tipo de regressão múltipla usado quando tanto a variável preditora quanto a variável resposta são multivariadas. Diferente da semelhante análise de correspondência canônica, a análise de redundância não faz nenhuma premissa acerca da forma dos eixos preditores ou resposta. [12]

análise de tabela de contingência (*contingency table analysis*) Os métodos estatísticos usados para analisar conjuntos de dados em que tanto a variável resposta quanto a preditora são categóricas. [7, 11]

análise de variância (ANOVA – *analysis of variance*) Um método de partição da soma dos quadrados, creditada a Fisher. É usada primariamente para testar a hipótese nula estatística de que tratamentos distintos não possuem efeitos sobre uma variável resposta medida. [5, 9, 10]

análise de variância multivariada (MANOVA – *multivariate analysis of variance*) Um método para testar a hipótese nula de que os vetores médios de múltiplos grupos são todos iguais uns aos outros. Uma ANOVA para dados multivariados. [12]

análise discriminante (*discriminant analysis*) Um método para classificar observações multivariadas quando os grupos foram definidos com antecedência. A análise discriminante também pode ser usada como um teste *post hoc* para detectar diferenças entre grupos quando análise de variância multivariada (MANOVA) tenha produzido resultados significativos. [12]

análise fatorial (*factor analysis*) Um método de ordenação que re-escala cada observação original como uma combinação linear de um número menor de parâmetros, chamados de fatores comuns. Foi desenvolvida no começo do século 20 como forma de calcular a "inteligência geral" a partir de uma bateria de testes com pontuação. [12]

análise paramétrica (*parametric analysis*) Uma das três principais estruturas de análises estatísticas. Ela recai sobre o requisito de que os dados foram tirados de uma variável aleatória com distribuições de probabilidade definidas por parâmetros fixos. Também é chamada de análise frequentista ou estatística frequentista. *Comparar com* **inferência bayesiana** e **análises de Monte Carlo.** [5]

ANCOVA *Ver* **análise de covariância.** [10]

ANOVA *Ver* **análise de variância.** [5, 9, 10]

ARIMA *Ver* **autorregressivo integrado de médias móveis.** [7]

arquivos simples (*flat files*) Arquivos de computador que contêm dados organizados em forma de tabela com apenas 2 dimensões (linhas e colunas). Cada dado é colocado em uma célula diferente. [8]

árvore de classificação (*classification tree*) Um modelo ou exibição visual de dados categóricos que é construído dividindo o conjunto de dados em grupos sucessivamente menores. A divisão continua até que nenhuma melhora no ajuste do modelo seja obtida. Uma árvore de classificação é análoga a uma chave taxonômica, exceto que as divisões em uma árvore de classificação refletem probabilidades e não identidades. *Ver também* **diagrama de agrupamento; impureza, árvore de regressão.** [11]

árvore de regressão (*regression tree*) Um modelo, ou representação visual, de dados contínuos que é construído por meio da divisão do conjunto de dados em grupos sucessivamente menores. A divisão continua até que nenhuma melhora no ajuste do modelo seja obtida. Uma árvore de regressão é análoga a uma chave taxonômica, exceto que as divisões em uma árvore de regressão refletem probabilidades, e não identidades. *Ver também* **árvore de classificação; diagrama de agrupamento.** [11]

árvore lógica (*logic tree*) Um exemplo do método hipotético-dedutivo. O investigado trabalho ao longo de uma árvore lógica escolhendo entre uma de duas alternativas a cada ponto de decisão. A árvore vai do geral (p. ex., animal ou vegetal) para o específico (p. ex., 4 dedos ou 3?). Quando não restar mais escolhas, a solução foi obtida. Uma chave dicotômica é um exemplo de árvore lógica. [4]

assimetria (*skewness*) A descrição de quão longe uma distribuição está de ser simétrica. Abreviada como g_1 e calculada a partir do terceiro momento central. *Ver também* **assimetria à direita** e **assimetria à esquerda**. [3]

assimetria à direita (*right-skewed*) Uma distribuição que tem a maioria de suas observações inferiores à média aritmética, mas uma longa cauda de observações superiores à média aritmética. A assimetria de uma distribuição com assimetria à direita é maior que 0. *Comparar com* **assimetria à esquerda**. [3]

assimetria à esquerda (*left-skewed*) Uma distribuição que tem a maioria de suas observações superiores à média aritmética, mas uma longa cauda de observações inferiores à média aritmética. A assimetria de uma distribuição com assimetria à esquerda é inferior a 0. *Comparar com* **assimetria à direita**. [3]

associação (*associative*) Em matemática, é a propriedade de que $(a + b) + c = a + (b + c)$. [A]

autorregressivo (*autoregressive*) Um modelo estatístico em que a variável resposta (Y) depende de um ou mais valores prévios. Muitos modelos do crescimento populacional são autorregressivos, como, por exemplo, aqueles nos quais o tamanho da população no tempo $t + 1$ é dependente do da população no tempo t. [6]

autorregressivo integrado de médias móveis (**ARIMA** – *autoregressive integrated moving average*) Um modelo estatístico usado para analisar dados de séries temporais que incorpora tanto a dependência temporal nos dados (*Ver* **autorregressivo**) quanto os parâmetros que estimam as mudanças devido aos tratamentos experimentais. [7]

autovalor (*eigenvalue*) A solução ou as soluções para um conjunto de equações lineares de forma $\mathbf{Ax} = \lambda\mathbf{x}$. Nesta equação, $\mathbf{A}$ é uma matriz quadrada e $\mathbf{x}$ é um vetor coluna. O autovalor é a escalar, ou um único número, λ. *Comparar com* **autovetor**. [12]

autovetor (*eigenvector*) O vetor coluna $\mathbf{x}$ que resolve o sistema de equações lineares $\mathbf{Ax} = \lambda\mathbf{x}$. *Comparar com* **autovalor**. [12]

axioma (*axiom*) Um princípio (matemático) estabelecido que é universalmente aceito, verdadeiro por autoevidência e não necessita de provas formais. [1]

base (*base*) A função logarítmica $\log_b a = X$ é descrita por dois parâmetros, a e b, tal que $b^X = a$ é equivalente a $\log_b a = X$. O parâmetro b é a base do logaritmo. [8]

beta (*beta*) A segunda letra (β) do alfabeto grego. Na estatística, é usada para representar a probabilidade aceitável de cometer um erro Tipo II. Ela também aparece, com subscritos, como símbolo para os coeficientes e parâmetros de regressão. [4]

BIC *Ver* **critério de informação bayesiano**. [9]

bi-plot (*bi-plot*) Um único gráfico dos resultados de duas ordenações relacionadas. Comumente usado para ilustrar os resultados da análise de correspondência, análise de correspondência canônica e análise de redundância. [12]

bloco (*block*) Uma área, ou um intervalo de tempo, dentro da qual fatores não manipulados experimentalmente são considerados homogêneos. [7]

bondade do ajuste (*goodness-of-fit*) Grau com que um conjunto de dados é acuradamente predito por uma dada distribuição ou modelo. [11]

bootstrap (*bootstrap*) Na língua inglesa, a palavra *bootstrap* significa "meios para conseguir algo com seu próprio esforço, sem ajuda de mais ninguém". Na estatística, o *bootstrap* é um procedimento de aleatorização (ou Monte Carlo), no qual as observações em um conjunto de dados são reamostradas com reposição a partir do conjunto de dados original, e o conjunto de dados reamostrado é então usado para recalcular a estatística teste de interesse. Isto é repetido um grande número de vezes (em geral 1.000 – 10.000, ou mais). A estatística teste que foi calculada com o conjunto de dados original é comparada com a distribuição daquela resultante das repetidas reamostragens dos dados, para obter um valor de P exato. *Ver também* **jackknife**. [5]

box plot (*box plot*) Uma ferramenta visual que ilustra a distribuição de um conjunto de dados univariado. Ilustra a mediana, os quartis superior e inferior, os decis superior e inferior, e qualquer discrepância (*outlier*). [8]

cálculo de probabilidade (*probability calculus*) A matemática usada para lidar com probabilidades. [1]

carga (*loadings*) Em uma análise de componentes principais, os parâmetros que são multiplicados por cada variável de uma observação multivariada para gerar uma nova variável, ou escore de componente principal para cada observação. Cada um desses escores é uma combinação linear das variáveis originais. [12]

carga dos fatores (*factor loadings*) Os coeficientes lineares com que os fatores comuns em uma análise fatorial são multiplicados para obter uma estimativa da variável original. [12]

casual (*haphazard*) Atribuição de indivíduos ou populações a grupos de tratamentos que não é verdadeiramente aleatória. *Comparar com* **randomização**. [6]

CCA *Ver* **análise de correspondência canônica**. [12]

célula (*cell*) Um campo em uma planilha. É identificada pelo número de sua linha e coluna e contém um único valor, ou dado. [8]

centroide (*centroid*) O vetor médio de uma variável multivariada. Pense nele como o centro de uma nuvem de pontos no espaço multivariado, ou o ponto em que você poderia equilibrar essa nuvem de pontos na ponta de uma agulha. [12]

circularidade (*circularity*) A premissa da análise de variância de que a variação entre amostras dentro de subparcelas ou blocos é a mesma ao longo de todas as subparcelas ou blocos. [7]

classificação (*classification*) O processo de posicionar, colocar em grupos. Um termo geral para os métodos multivariados, como a análise de agrupamento ou a análise discriminante, que estão preocupados em agrupar observações. *Comparar com* **ordenação**. [12]

coeficiente binomial (*binomial coefficient*) A constante pela qual multiplicamos a probabilidade de uma variável aleatória de Bernoulli para obter uma variável aleatória binomial. O coeficiente binomial ajusta a probabilidade para levar em conta múltiplas combinações equivalentes de cada um dos dois resultados possíveis. Ele é escrito como:

$$\binom{n}{X}$$

e é lido como "n sobre X". É calculado como:

$$\frac{n!}{X!(n-X)!}$$

onde "!" indica o fatorial. [2]

coeficiente de caminhos (*path coefficient*) Um tipo de parâmetro de regressão parcial calculado na análise de caminhos que indica a magnitude e o sinal de efeito de uma variável sobre outra. [9]

coeficiente de correlação produto-momento (*r*) (*product-moment correlation coefficient*) Uma medida de quão bem duas variáveis são relacionadas uma com a outra. Ele pode ser calculado como a raiz quadrada do coeficiente de determinação, ou como a soma dos produtos cruzados dividida pela raiz quadrada do produto da soma dos quadrados das variáveis X e Y. [9]

coeficiente de determinação (r^2) (*coefficient of determination*) A quantidade de variabilidade da variável resposta, que é explicada por um modelo de regressão linear simples. É igual à soma dos quadrados da regressão dividida pela soma dos quadrados total. [9]

coeficiente de dispersão (*coefficient of dispersion*) A quantidade obtida pela divisão da variância da amostra pela média da amostra. O coeficiente de dispersão é usado para determinar se os indivíduos distribuídos com padrão espacial agregado, regular, aleatório ou hiperdisperso. [3]

coeficiente de variação (CV) (*coefficient of variation*) A quantidade obtida pela divisão do desvio padrão da amostra pela média da amostra. O coeficiente de variação é útil para comparar a variabilidade entre diferentes populações, pois foi ajustada (padronizada) para a média da população. [3]

colinearidade (*collinearity*) A correlação entre duas variáveis preditoras. *Ver também* **multicolinearidade**. [7, 9]

combinação (*combination*) A organização de n objetos discretos com X objetos a cada vez. O número de combinações de um grupo de n objetos é calculado usando o coeficiente binomial. *Comparar com* **permutação**. [2]

combinação linear (*linear combination*) A re-expressão de uma variável como a combinação de outras, cujos parâmetros da nova expressão são lineares. Por exemplo, a equação $Y = b_0X_0 + b_1X_1 + b_2X_2$ expressa a variável Y como uma combinação linear de X_0 a X_2, pois os parâmetros b_0, b_1 e b_2 são lineares. Em contraste, a equação $Y = a_0X_0 + X_1^{a_1}$ não é uma combinação linear, pois o parâmetro a_1 é exponencial, e não linear. Note que a segunda equação pode ser transformada em uma combinação linear, tirando os logaritmos de ambos lados da equação. Nem sempre será o caso, entretanto, que combinações não lineares possam ser transformadas em lineares. [12]

complemento (*complement*) Na teoria de conjuntos, o complemento A^c de um conjunto A é tudo que não está incluído no conjunto A. [1]

completamente cruzados (*fully crossed*) Um delineamento de análise de variância multifatorial em que todos os níveis de dois ou mais tratamentos são testados ao mesmo tempo em um único experimento. [7, 10]

completo (*exhaustive*) Um conjunto de resultados é dito completo se inclui todos os resultados possíveis de um evento. *Comparar com* **exclusivo**. [1]

componente de variação (*component of variation*) Quanto cada fator, em um modelo de análise de variância, explica da variabilidade dos dados observados. Inicialmente, a variação total nos dados pode ser fracionada em variação entre grupos e em variação dentro de grupos. Em alguns delineamentos de ANOVA; a variação pode ser fracionada em componentes adicionais. [10]

componentes principais (*principal components*) Os eixos principais ordenados pela quantidade de variabilidade que eles carregam do conjunto original de dados. [12]

comutativo (*commutative*) Em matemática, é a propriedade de que $a + b = b + a$. [A]

confundido (*confounded*) A consequência de uma variável medida estar associada a outra variável que pode não ter sido medida ou controlada. [4, 6]

conjunto (*set*) Uma coleção de objetos discretos, consequências ou resultados. Os conjuntos são manipulados usando as operações de **união, interseção** e **complemento**. [1]

conjunto vazio Um conjunto sem membros ou elementos. [1]

conjuntos discretos (*discrete sets*) Um conjunto de resultados discretos. [1]

contraste (*contrast*) A comparação entre grupos de tratamentos em um delineamento de ANOVA. Os contrastes se referem às comparações decididas *antes* de um experimento ser realizado. [10]

correlação (*correlation*) O método de análises usado para explorar relações entre duas variáveis. Na análise de correlação, nenhuma hipótese sobre relação de causa e efeito é criada. *Comparar com* **regressão**. [7, 9]

correlativo (*correlative*) Um tipo de relações que é observado, mas que ainda não foi investigado experimentalmente de forma controlada. A existência de dados correlativos leva à expressão "correlação não implica causa". [4]

covariável (*covariate*) Uma variável contínua que é medida em cada réplica em um delineamento de análise de covariância. A covariável é usada para explicar a variação residual nos dados e aumentar o poder do teste em detectar diferenças entre tratamentos. [10]

covariância (*covariance*) A soma do produto das diferenças entre cada observação e seu valor esperado, dividido pelo tamanho amostral. Se a covariância não for dividida pelo tamanho amostral, é chamada de **soma dos produtos cruzados**. [9]

critério de informação (*information criterion*) Qualquer método para selecionar entre modelos estatísticos que leve em conta o número de parâmetros nos diferentes modelos em geral qualificados (como no critério de informação de Akaike ou critério de informação bayesiano). [9]

critério de informação bayesiano (**BIC** – *Bayesian information criterion*) Um método para comparar as distribuições de probabilidades *a posteriori* de dois modelos alternativos quando a de probabilidades *a priori* não é informativa. Ele desconta o fator de Bayes pelo número de parâmetros em cada um dos modelos. Como com o critério de informação de Akaike, esse modelo é considerado o melhor. [9]

critério de informação de Akaike (**AIC** – Akaike's *information criterion*) Uma medida do poder explicativo de um modelo estatístico que leva em conta o número de parâmetros no modelo. Ao comparar muitos modelos para um mesmo fenômeno (os modelos devem compartilhar no mínimo alguns parâmetros), o modelo com o menor valor do critério de informação de Akaike é considerado o melhor. Usado comumente para seleção de modelos em regressão e análise de caminhos. [9]

curtose (*kurtosis*) Uma medida de quão agregada ou dispersa uma distribuição é relativa a seu centro. É calculada usando o quarto momento central e é simbolizada por g_2. *Ver também* **platicúrtica** e **leptocúrtica**. [3]

CV *Ver* **coeficiente de variação**. [3]

---

dados bivariados (*bivariate data*) Dados que consistem de duas variáveis para cada observação. Em geral são ilustrados em gráficos de dispersão. [8]

dados multivariados (*multivariate data*) Os dados consistem de mais de duas variáveis para cada observação. Com frequência são ilustrados usando matrizes de gráficos de dispersão ou outras ferramentas de visualização em mais dimensões. [8, 12]

dados univariados (*univariate data*) Os dados consistem de uma única medida para cada observação. Muitas vezes são ilustrados usando *box plots*, histogramas ou gráficos de caule-e-folhas. [8, 12]

decil (*decile*) Um décimo. Em geral refere-se aos decis superior e inferior – os 10% superiores e os 10% inferiores – de uma distribuição. [3]

decomposição (*decomposition*) Em álgebra de matrizes, é a re-expressão de qualquer matriz $m \times n$ A

conforme o produto de três matrizes, $\mathbf{VWU}^{\mathrm{T}}$, onde **V** é uma matriz com as mesmas dimensões de **A**, **U** e **W** são matrizes quadradas (dimensão $n \times n$); **V** e **U** são matrizes de colunas ortonormais; e **W** é uma matriz diagonal com valores singulares na diagonal. [A]

dedução (*deduction*) O processo de raciocínio científico que progride do geral (todos os cisnes são brancos) para o específico (esta ave, sendo um cisne, é branca). *Comparar com* **indução**. [4]

delineamento ADCI (*BACI design*) Um método experimental para avaliar os efeitos de um tratamento (frequentemente um impacto ambiental adverso) em relação a uma população sem ameaças. Nos delineamentos ADCI (Antes-Depois-Controle-Impacto) com frequência há pouca replicação espacial, portanto a replicação é temporal. O efeito nas populações controle e nas populações ameaçadas são medidos tanto antes quanto depois do tratamento. [6, 7]

delineamento aditivo (*additive design*) Um delineamento experimental usado para estudar competição entre uma espécie de interesse alvo, ou focal, e seus competidores. Em um delineamento aditivo, a densidade da espécie-alvo é mantida constante e os diferentes tratamentos representam diferentes números de competidores. *Comparar com* **delineamentos substitutivos** e de **superfície de resposta**. [7]

delineamento aninhado (*nested design*) Qualquer delineamento de análise de variância em que há subamostragem dentro das réplicas. [7, 10]

delineamento com parcelas subdivididas (*split-plot design*) Um delineamento de análise de variância multifatorial em que cada parcela experimental é subdividida em subparcelas, e cada uma recebe um tratamento diferente. [7, 10]

delineamento com um fator (*single factor design*) Delineamento de análise de variância que manipula apenas uma variável. *Comparar com* **delineamento multifatorial**. [7]

delineamento de análise de variância (*analysis of variance designs*) A classe de configurações experimentais usada para explorar relações entre variáveis preditoras categóricas e variáveis resposta contínuas. [7]

delineamento de blocos aleatorizados (*randomized block design*) Um delineamento de análise de variância em que conjuntos de réplicas, e todos os níveis dos tratamentos, são dispostos em uma área física no espaço ou em um ponto fixo no tempo (um bloco), no intuito de reduzir a variabilidade nas condições não manipuladas. [7, 10]

delineamento de dois fatores (*two-way design*) Um delineamento de ANOVA com dois efeitos principais em que cada um deles tem dois ou mais níveis de tratamento. [7, 10]

delineamento de medidas repetidas (*repeated measures design*) Um tipo de ANOVA em que múltiplas observações são feitas em uma mesma réplica individual, em diferentes pontos no tempo. [7, 10]

delineamento de regressão (*regression design*) A classe de esboço experimental usado para explorar relações entre variáveis preditoras contínuas e variáveis resposta contínuas ou categóricas. [7]

delineamento de superfície de resposta (*response surface design*) Um delineamento experimental usado para estudar competição entre uma espécie de interesse alvo, ou focal, e seus competidores. Em um delineamento de superfície de resposta, a densidade da espécie-alvo e de seus competidores são variadas sistematicamente. Os delineamentos de superfície de resposta também podem ser usados em vez de ANOVA quando os fatores são melhor representados como variáveis contínuas. *Comparar com* **delineamento aditivo** e **delineamento substitutivo**. [7]

delineamento fatorial (*factorial design*) Um delineamento para análise de variância que inclui todos os níveis dos tratamentos (ou fatores) de interesse. [7]

delineamento multifatorial (*multi-factor design*) Delineamento de análise de variância que inclui dois ou mais tratamentos ou fatores. [7]

delineamento substitutivo (*substitutive design*) Um delineamento experimental usado para estudar a competição entre uma espécie de interesse-alvo, ou focal, e seus competidores. Em um delineamento substitutivo, a densidade total da espécie-alvo e de seus competidores é mantida constante em certo nível, mas a proporção relativa da espécie-alvo e seus competidores é variada sistematicamente. *Comparar com* **delineamento aditivo** e com **delineamento de superfície de resposta**. [7]

dendrograma (*dendrogram*) Um diagrama (*grama*) de árvore (*dendro*). Um método de representar graficamente os resultados de uma análise de agrupamentos. Em princípio, é similar a uma árvore de classificação. Também é muito usado para ilustrar a relação entre espécies (filogenias). [12]

desigualdade triangular (*triangle inequality*) Para três objetos multivariados, a, b e c, a distância entre a e b mais a entre b e c sempre é maior que, ou igual à, distância entre a e c. [12]

destendenciada (*unbiased*) A propriedade estatística de medidas, observações ou valores não serem

nem consistentemente muito grandes nem muito pequenas. *Ver também* **acurácia e precisão**. [2]

desvio (*deviation*) Diferença entre a predição do modelo e dos valores observados. [6]

desvio padrão (DP) (*standard deviation – sd*) Uma medida de dispersão. Igual à raiz quadrada da variância. [3]

desvio padrão amostral (*sample standard deviation*) Uma estimativa, sem viés, do desvio padrão de uma amostragem. Igual à raiz quadrada da variância amostral. [3]

determinante (*determinant*) Em álgebra de matrizes, é a escalar calculada como a soma de todos os produtos possíveis, contendo um elemento na linha e um na coluna de uma dada matriz. [A]

diagrama de Venn (*Venn diagram*) Um gráfico usado para ilustrar as operações de união, intercessão e complemento de conjuntos. [11]

dimensão (*dimension*) Um vetor bidimensional especificando o número de linhas e o de colunas em uma matriz. [A]

discrepâncias (*outliers*) Dados inesperadamente grandes ou pequenos. [3, 8]

dispersão (*spread*) Uma medida da variação dentro de um conjunto. [3]

dissimilaridade (*dissimilarity*) Quão diferentes dois objetos são. Normalmente quantificados com uma **distância**. [12]

distância (*distance*) Quão separados dois objetos estão. Há muitas formas para medir distâncias, a mais familiar é a **distância euclidiana**. [12]

distância euclidiana (*Euclidean distance*) Uma medida de distância ou dissimilaridade entre dois objetos. Se cada objeto pode ser descrito por um conjunto de coordenadas em um espaço n-dimensional, a distância euclidiana pode ser calculada como:

$$d_{i,j} = \sqrt{\sum_{k=1}^{n} (y_{i,k} - y_{j,k})^2} \qquad [12]$$

distribuição de probabilidades (*probability distribution*) A distribuição de resultados X_i de uma variável aleatória X. Para variáveis aleatórias contínuas, a distribuição de probabilidades é uma curva suave e um histograma apenas é uma aproximação da distribuição de probabilidades. [2]

distribuição de probabilidades *a posteriori* (*posterior probability distribution*) A distribuição resultante da aplicação do Teorema de Bayes a um determinado conjunto de dados. Essa distribuição dá a probabilidade para o valor de qualquer hipótese de interesse conforme os dados observados. [5, 1]

distribuição de probabilidades *a priori* (*prior probability distribution*) Uma de duas distribuições (a outra é a verossimilhança) usada no numerador do Teorema de Bayes para calcular a distribuição de probabilidades *a posteriori*. Ela define a probabilidade para o valor de qualquer hipótese de interesse antes que os dados sejam coletados. *Ver também* **probabilidade *a priori* e verossimilhança**. [5, 9]

distribuição-F (*F-distribution*) A distribuição esperada de uma variável aleatória F, calculada como uma variância (em geral a soma dos quadrados entre grupos ou $SQ_{entre\ grupos}$) dividida por outra (geralmente a soma dos quadrados dentro de grupos, ou $SQ_{dentro\ de\ grupos}$). [5]

distribuição normal padrão (*standard normal distribution*) Uma distribuição normal de probabilidades com média = 0 e variância = 1. [2]

distribuição-*t* (*t-distribution*) Uma modificação da distribuição de probabilidades normal padrão. Para tamanhos amostrais pequenos, a distribuição-*t* é leptocúrtica, mas, conforme o tamanho amostral aumenta, a distribuição-*t* se aproxima de uma distribuição normal padrão. As probabilidades de cauda da distribuição-*t* são usadas para calcular intervalos de confiança. [3]

distributivo (*distributive*) Em matemática, é a propriedade de que $c(a + b) = ca + cb$. [A]

dobradiça (*hinge*) A posição dos quartis superior e inferior em um gráfico de caule-e-folha. [8]

DP *Ver* **desvio padrão**. [3]

EDA *Ver* **análise de dados exploratória usando gráficos**. [8]

efeito do arco (*horseshoe effect*) Uma propriedade indesejável de muitos métodos de ordenação que acabam exagerando a média e comprimindo os extremos de um gradiente ambiental. Ele leva a um gráfico de ordenação que é curvo, em forma arco, ferradura, ou círculos. [12]

efeito principal (*main effect*) O efeito individual de uma única variável. Na ausência de qualquer interação os efeitos principais são aditivos: a resposta de um experimento pode ser acuradamente predita conhecendo-se os efeitos principais de cada fator individual. *Comparar com* **interação**. [7, 10]

efeitos aleatórios (*random effects*) Em um delineamento de análise de variância, o conjunto de níveis de tratamento que representam uma variável, um subconjunto aleatório de todos os níveis de tratamento possíveis. Também conhecido como fator aleatório. *Comparar com* **efeitos fixos**. [7, 10]

efeitos fixos (*fixed effects*) Em um delineamento de análise de variância, o conjunto de níveis dos tratamentos que representa todos os níveis de tratamentos possíveis. Também conhecido como fator fixo. *Comparar com* **efeitos aleatórios**. [7, 10]

eixo maior (*major axis*) Em uma análise de componentes principais, é o primeiro eixo principal. Além disso, é o eixo mais longo de uma elipse. [12]

eixo principal (*principal axii*) Uma nova variável criada por uma ordenação em que as variáveis originais são ordenadas. [12]

eixo-*x* (*x-axis*) A linha horizontal em um gráfico, também chamado de abscissa. O eixo-*x* em geral representa a variável causal, independente ou preditora em um modelo estatístico. [1]

eixo-*y* (*y-axis*) A linha vertical em um gráfico, também chamado de ordenada. O eixo-*y* geralmente representa o resultado ou variável dependente ou resposta em um modelo estatístico. [1]

elemento (*element*) Um valor em uma matriz. Na maioria das vezes é indicado com a_{ij}, onde i e j são os números da linha e da coluna, respectivamente, da matriz. [12]

eliminação regressiva (*backward elimination*) Um método de regressão gradativa em que você começa com um modelo contendo todos os parâmetros possíveis e os elimina um por vez. *Ver também* **regressão gradativa** (*stepwise regression*); *comparar com* **seleção progressiva** (*forward selection*). [9]

erro (*error*) Um valor que não representa a medida ou observação original. Ele pode ser causado por desatenção no campo, instrumentos com mau funcionamento ou erros de digitação cometidos ao transferir os dados do caderno de campo ou folhas de dados para as planilhas. [8]

erro-padrão da média (*standard error of the mean*) Uma estimativa assintótica do desvio padrão da população. Com frequência apresentada em publicações em vez do desvio padrão amostral. O erro padrão da média é menor que o amostral e pode dar a falsa impressão de menos variabilidade nos dados. [3]

erro-padrão da regressão (*standar error of regression*) Para um modelo de regressão, a raiz quadrada da soma dos quadrados dos resíduos dividida pelos graus de liberdade do modelo. [9]

erro Tipo I (*Type I error*) Rejeição errônea de uma hipótese nula estatística verdadeira. Em geral, é indicado pela letra grega alfa (α). [4]

erro Tipo II (*Type II error*) Aceitar incorretamente uma hipótese nula estatística que é falsa. Em geral, indicado pela letra grega beta (β). O poder estatístico é igual a $1 - \beta$. [4]

escalar (*scalar*) Um único número, não uma matriz. [A]

escalonamento multidimensional não métrico (NMDS – *non-metric multidimensional scaling*) Um método de ordenação que preserva a ordem da classificação das distâncias, ou dissimilaridades, originais entre as observações. Este método geral e robusto de ordenação pode usar qualquer tipo de medida de distância ou dissimilaridade. [12]

escore do componente principal (*principal component score*) O valor de cada observação, que é uma combinação linear das variáveis originais medidas para aquela observação. É determinado por uma análise de componentes principais. [12]

escores ajustados dos locais (*fitted site scores*) Um resultado da análise de redundância. Os escores dos locais ajustados são a matriz **Z**, que resulta de uma ordenação dos dados originais em relação às características ambientais

escores das espécies (*species scores*) Um resultado da análise de redundância. Os escores das espécies são a matriz de autovetores **A**, que resulta da análise de componentes principais da matriz padronizada de variância-covariância dos valores preditos da variável resposta ($\hat{\mathbf{Y}}$). É usada para gerar tanto os escores dos locais quanto os escores ajustados dos locais. [12]

escores dos locais (*site scores*) Um resultado da análise de redundância. Os escores dos locais são a matriz **F** resultante de uma ordenação dos dados originais em relação a características ambientais $\mathbf{F} = \mathbf{YA}$, onde **Y** é a resposta multivariada e **A** é a matriz de autovetores que resulta da análise de componentes principais dos valores esperados (ajustados) da variável resposta. *Comparar com* **escores ajustados dos locais** e **escores das espécies**. [12]

$$\mathbf{Z} = \hat{\mathbf{Y}}\mathbf{A}$$

onde $\hat{\mathbf{Y}}$ são os valores preditos da variável resposta multivariada e **A** é a matriz de autovetores, que resulta de uma análise de componentes principais dos valores esperados (ajustados) da variável resposta. *Comparar com* **escores dos locais** e **escores das espécies**. [12]

escore-*Z* (*Z-score*) O resultado de transformar uma variável pela subtração de sua média amostral e dividindo a diferença pelo desvio padrão amostral. [12]

esfericidade (*sphericity*) A premissa dos métodos de teste de hipóteses multivariado de que as covariâncias de cada grupo são iguais umas às outras. [12]

espaço amostral (*sample space*) O conjunto de todos os resultados possíveis ou resultados de um evento ou experimento. [1]

esquema de pares casados (*matched-pairs layout*) Uma forma de delineamento em blocos aleatorizados na qual as réplicas são selecionadas para serem similares em tamanho corporal, ou outra característica, e então são atribuídas de maneira aleatória aos tratamentos. [7]

estatística assintótica (*asymptotic statistics*) Estimadores estatísticos que são calculados com base na premissa de que, se o experimento for repetido infinitas vezes, o estimador poderia convergir no valor calculado. *Ver também* **inferência frequentista**. [3]

estatística descritiva (*summary statistics*) Números que descrevem a posição e a dispersão de dados. [3]

estatística-F (*F-statistic*) O resultado, em uma análise de variância (ANOVA) ou em regressão, da divisão da soma dos quadrados entre grupos, ($SQ_{entre\ grupos}$) pela soma dos quadrados dentro de grupos ($SQ_{dentro\ de\ grupos}$). Ela tem dois graus de liberdade diferentes – os do numerador, associados com a $SQ_{entre\ grupos}$ e os graus de liberdade do denominador são associados com a $SQ_{dentro\ de\ grupos}$. Esses dois valores também definem a forma da distribuição-F. Para determinar sua significância estatística, ou valor de *P*, a estatística-F é comparada ao valor crítico da distribuição-F, dado o tamanho amostral, a probabilidade de erro Tipo I e os graus de liberdade do numerador e do denominador. [5]

estatística não paramétrica (*non-parametric statistics*) Um ramo da análise estatística que não depende dos dados terem sido tirados de uma variável aleatória definida e com distribuição de probabilidade conhecida. É frequentemente substituída por análises de Monte Carlo. [5]

estatística do teste (*test statistic*) O resultado numérico da manipulação de dados que são usados para examinar hipóteses estatísticas. As estatísticas do teste são comparadas a valores que poderiam ser obtidas de cálculos idênticos se a hipótese nula estatística de fato fosse verdadeira. [4]

estimador destendenciado (*unbiased estimator*) Um parâmetro estatístico calculado, que não é consistentemente maior nem menor que o valor real. [3, 9]

estimadores-*M* (*M-estimators*) Um método de regressão para minimizar o efeito de dados destoantes. Eles calculam inclinações de regressões após dar menos peso às maiores somas dos quadrados. [9]

estresse (*stress*) Uma medida de quão bem os resultados de um escalonamento multidimensional não métrico prediz os valores observados. Além disso, o resultado de gastar muito tempo em análises estatísticas e não tempo suficiente no campo. [12]

evento (*event*) Uma simples observação ou processo com começo e fim claramente definidos. [1]

evento complexo (*complex event*) Na teoria de conjuntos, eventos complexos são coleções de eventos simples. Se *A* e *B* são dois eventos independentes, então $C = A\ ou\ B$, expresso como $C = A \cup B$ e lê-se "*A* união com *B*", será um evento complexo. Se *A* e *B* forem eventos independentes, com probabilidades associadas P_A e P_B, então a probabilidade de *C* é $P_C = P_A + P_B$. *Comparar com* **eventos compartilhados**. [1]

eventos compartilhados (*complex event*) Na teoria de conjuntos, os eventos compartilhados são coleções de eventos simples. Se *A* e *B* são dois eventos independentes, então $C = A\ e\ B$, escrito como $C = A \cap B$ e lido como "*A* intercessão com *B*", é um evento compartilhado. Se *A* e *B* são dois eventos independentes, cada um associado a uma probabilidade P_A e P_B, então a probabilidade de *C* (P_C) é igual a $P_A \times P_B$. *Comparar com* **evento complexo**. [1]

exclusivo (*exclusive*) Um conjunto de resultados é dito exclusivo se não houver nenhuma sobreposição em seus valores. *Comparar com* **completo**. [1]

expectativa (*expectation*) A média aritmética, média ou valor esperado de uma distribuição de probabilidades. [1, 2]

experimento (*experiment*) Um conjunto de observações replicadas, em geral realizadas sob condições controladas. [1]

experimento de pressão (*press experiment*) Um experimento manipulativo no qual um tratamento é aplicado e reaplicado ao longo da duração do experimento com objetivo de manter a força do tratamento. *Comparar com* **experimento de pulso**. [6]

experimento de pulso (*pulse experiment*) Um experimento manipulativo em que o tratamento é aplicado apenas uma vez, e então a réplica "tratada" é deixada para se recuperar da manipulação. *Comparar com* **experimento de pressão**. [6]

experimento de trajetória (*trajectory experiment*) Um tipo de experimento natural em que todas as réplicas são amostradas repetidamente em muitos pontos no tempo. A replicação é baseada na variabilidade temporal. Os experimentos de trajetória frequentemente são analisados usando séries temporais ou modelos autorregressivos. *Comparar com* experimento **fotográfico**. [6]

experimento fotográfico (*snapshot experiment*) Um tipo de experimento natural em que todas as réplicas são amostradas em um ponto no tempo. A re-

plicação é fundamentada na variabilidade espacial. *Comparar com* **experimento de trajetória.** [6]

experimento manipulativo (*manipulative experiment*) Aquele no qual o investigador deliberadamente aplica um ou mais tratamentos a uma população amostral, ou populações, e então observa o resultado do tratamento. *Comparar com* **experimento natural.** [6]

experimento natural (*natural experiment*) A comparação entre dois ou mais grupos que não foram manipulados de nenhuma forma pelo investigador; pelo contrário, o delineamento recai sobre a variação natural entre grupos. Os experimentos naturais idealmente comparam grupos que são iguais em todas as formas, exceto para um único fator causal de interesse. Embora os experimentos naturais possam ser analisados com muitas das mesmas estatísticas que os manipulativos, as inferências resultantes podem ser mais fracas por causa de variáveis confundidas que não foram controladas. [1, 6]

extensão (*extent*) A área espacial total ou amplitude temporal que é coberta por todas as unidades amostrais em um estudo experimental. *Comparar com* **grão.** [6]

extrapolação (*extrapolation*) A estimativa de valores não observados fora da gama dos valores observados. Em geral, é menos confiável que a **interpolação.** [9]

---

fator (*factor*) Em delineamentos de ANOVA, um conjunto de níveis de um único tratamento experimental. [7]

fator de Bayes (*Bayes' factor*) A probabilidade *a posteriori* em relação da hipótese nula com a alternativa. Se as probabilidades *a priori* das duas hipóteses forem iguais, ele pode ser calculado como a razão das verossimilhanças das duas hipóteses. [9]

fator de parcela completa (*whole plot factor*) Em um delineamento de parcelas subdivididas, o tratamento que é aplicado a um bloco ou parcela inteira. *Comparar com* **fator de subparcela.** [7]

fator de subparcela (*Subplot factor*) Em um delineamento de parcelas subdivididas, é o tratamento que é aplicado dentro de um único bloco ou parcela. *Comparar com* **fator de parcela completa.** [7]

fator entre-sujeitos (*between-subjects factor*) Na ANOVA, o termo que leva em conta a variação entre grupos de tratamentos. Em um delineamento de medidas repetidas, o fator entre-sujeitos é equivalente ao de parcela completa de um delineamento de parcelas subdivididas (*split-plot design*). [7]

fator intrassujeitos (*within-subjects factor*) Em uma ANOVA, é o termo que explica a variância dentro dos grupos de tratamentos. Em um delineamento de medidas repetidas, o fator intrassujeitos é equivalente ao de subparcelas do delineamento de parcelas subdivididas. [10]

fatores comuns (*common factors*) Os parâmetros gerados em uma análise fatorial. Cada variável original pode ser re-expressa como uma combinação linear dos fatores comuns, mais um fator adicional que é único àquela variável. *Ver também* **análise fatorial** e **análise de componentes principais.** [12]

fatorial (*factorial*) A operação matemática indicada por um ponto de exclamação(!). Para um dado valor de n, o valor do $n!$, ou "n fatorial", é calculado como:

$$n \times (n-1) \times (n-2) \times \ldots \times (3) \times (2) \times (1)$$

Portanto, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$. Por definição, $0! = 1$, e fatoriais não podem ser calculados para números negativos e que não sejam inteiros. [2]

FDC *Ver* **função de distribuição cumulativa.** [2]

FDP *Ver* **função de densidade de probabilidade.** [2]

fechamento (*closure*) Uma propriedade matemática dos conjuntos. Os conjuntos são considerados fechados para uma operação matemática em particular, por exemplo a adição ou a multiplicação, se o resultado de aplicar uma operação a um membro ou a membros do conjunto também for um membro do conjunto. Por exemplo, o conjunto de números pares $\{2, 4, \ldots\}$ é fechado para a multiplicação, pois o produto de dois (ou mais) números inteiros também será um número par. Em contraste, o conjunto dos números naturais $\{0, 1, 2, \ldots\}$ não é fechado para a subtração, pois é possível subtrair um número natural de outro e obter um número negativo (p. ex., $3 - 5 = -2$), e números negativos não são números naturais. [1]

forma quadrática (*quadratic form*) Uma matriz 1×1 igual ao produto de $\mathbf{x}^T\mathbf{Ax}$, onde $\mathbf{x}$ é uma matriz coluna $n \times 1$ e $\mathbf{A}$ é uma matriz $n \times n$. [A]

frequência (*frequency*) O número de observações em uma categoria em particular. Os valores do eixo-y (ordenada) de um histograma são frequências de observações em cada grupo identificadas no eixo-x (abscissa). [1]

frequência relativa (*relative freauency*) *Ver* **proporção.** [11]

função (*function*) Uma regra matemática para atribuir um valor numérico a outro valor numérico. Em geral, as funções envolvem a aplicação de operações matemáticas (como adição ou multiplicação) ao valor inicial, para obter o resultado da

função. As funções muitas vezes são escritas com letras em itálico: a função

$$f(x) = \beta_0 + \beta_1 x$$

significa pegar o x, multiplicá-lo por β_1 e somar o β_0 ao produto. Isto dá um novo valor, que é relacionado ao valor original, x, pela operação da função $f(x)$. [2]

função densidade de probabilidade (FDP) (*probability density function*) A função matemática que atribui a probabilidade de observar um resultado X_i a cada valor de uma variável aleatória X. [2]

função de distribuição cumulativa (FDC) (*cumulative distribution function*) A área, acima de um ponto X, sob a curva que descreve uma variável aleatória. Formalmente, se X for uma variável aleatória com função de distribuição de probabilidade $f(X)$, então a função de distribuição cumulativa $F(X)$ é igual a:

$$P(x < X) = \int_{-\infty}^{X} f(x)dx$$

A função de distribuição cumulativa é usada para calcular a probabilidade em cauda de uma hipótese. [2]

função de influência (*influence function*) Um gráfico de diagnóstico para avaliar o impacto de cada observação na inclinação e no intercepto de um modelo de regressão. Neste gráfico, os interceptos obtidos por jackknife (no eixo-y) são plotados contra as inclinações obtidas por jackknife (no eixo-x). A inclinação desse gráfico sempre deve ser negativa, mas dados discrepantes em observações com efeitos indevidos (ou influência) no modelo de regressão. [9]

função log da verossimilhança (*log-likelihood function*) O logaritmo da função de verossimilhança. Maximizá-la é uma forma de determinar quais parâmetros de uma fórmula, ou hipótese, melhor explicam os dados observados. *Ver também* **verossimilhança**. [8]

função monótona contínua (*continuous monotonic function*) Uma função que é definida para todos valores em um intervalo e que preserva a ordem dos operandos. [8]

---

garantia de qualidade e controle de qualidade (GQ/CQ) (*quality assurance and quality control – QA/QC*) O processo pelo qual os dados (ou qualquer atividade) são monitorados para determinar que eles têm acurácia e precisão. [8]

gráfico de caule-e-folha (*stem-and-leaf plot*) Um histograma "rico em informação" em que o valor real de cada dado é ilustrado. [8]

gráfico de dispersão (*scatterplot*) Um gráfico de duas dimensões usado para ilustrar dados bivariados. A variável preditora, ou independente, geralmente é colocada no eixo-x e a variável resposta, ou dependente, no eixo-y. Cada ponto representa as medições dessas duas variáveis em cada observação. [8]

gráfico de resíduos (*residual plot*) Um gráfico de diagnóstico usado para determinar se as premissas da regressão ou ANOVA foram cumpridas. Mais comumente, os resíduos são plotados contra os valores preditos. [9]

gráfico mosaico (*mosaic plot*) Um gráfico destinado à ilustração de tabelas de contingência. O tamanho dos retângulos do mosaico é proporcional à frequência relativa de cada observação. [11]

grão (*grain*) O tamanho espacial ou a extensão temporal da menor unidade amostral em um estudo experimental. *Comparar com* **extensão**. [6]

graus de liberdade (gl) (*degrees of freedom – df*) Quantas observações independentes temos para uma certa variável que podem ser utilizadas para estimar um parâmetro estatístico. [3]

grupo (*group*) Um conjunto que possui quatro propriedades. Primeiro, o conjunto é fechado sob alguma operação matemática $\oplus$ (como adição ou multiplicação). Segundo, a operação é associativa, de modo que quaisquer três elementos a, b, c do conjunto $a \oplus (b \oplus c) = (a \oplus b) \oplus c$ (onde $\oplus$ indica qualquer operação matemática). Terceiro, o conjunto tem um elemento identidade I tal que para qualquer elemento a do conjunto $a \oplus I = a$. Quarto, o conjunto tem inversos, de maneira que para cada elemento a do conjunto existe outro elemento b para o qual $ab = ba = I$. *Ver também* **fechamento**. [1]

---

heterocedástico (*heteroscedastic*) A propriedade de um conjunto de dados que diz que a variância residual de todos os grupos de tratamentos são diferentes. *Comparar com* **homocedástico**. [8, 9]

hipótese (*hypothesis*) Uma afirmação testável de causa e efeito. *Ver também* **hipótese alternativa** e **hipótese nula**. [4]

hipótese alternativa (*alternative hypothesis*) A hipótese de interesse. Ela é a hipótese estatística que corresponde à questão científica que está sendo examinada com os dados. Diferente da hipótese nula, a hipótese alternativa diz que "algo está acontecendo." *Comparar com* **hipótese nula** e com **hipótese estatística**. [4]

hipótese científica (*scientific hypothesis*) Uma declaração de causa e efeito. A acurácia de uma hipótese científica é inferida a partir dos resultados de testar hipóteses estatísticas nulas e alternativas. [4]

hipótese extrínseca (*extrinsic hypothesis*) Um modelo estatístico em que os parâmetros são estimados a partir de outras fontes que não os dados. *Comparar com* **hipótese intrínseca**. [11]

hipótese intrínseca (*intrinsic hypothesis*) Um modelo estatístico cujos parâmetros são estimados a partir dos próprios dados. *Comparar com* **hipótese extrínseca**. [11]

hipótese nula (*null hypothesis*) A hipótese mais simples possível. Na ecologia e nas ciências ambientais, a hipótese nula geralmente é que a variabilidade observada nos dados pode ser atribuída inteiramente à aleatoriedade ou a erros de medida. *Comparar com* **hipótese alternativa**. *Ver também* **hipótese nula estatística**. [1, 4]

hipótese nula estatística (*statistical null hypothesis*) A hipótese formal de que não existe relação matemática entre duas variáveis, ou a forma estatística da hipótese científica de que qualquer variabilidade observada nos dados pode ser atribuída inteiramente à aleatoriedade ou a erro de medidas. *Ver também* **hipótese nula**. [4]

histograma (*histogram*) Um gráfico de barras que representa a distribuição de frequências. A altura de cada barra é igual à frequência da observação identificada no eixo-*x*. [1, 2]

homocedástico (*homocedastic*) A propriedade de um conjunto de dados cuja variância residual de todos os tratamentos é igual. *Comparar com* **heterocedástico**. [8]

IID Representa "independente e identicamente distribuído." Bons delineamentos experimentais resultam em réplicas e distribuição de erros que são IID. *Ver também* **ruído branco**, **independente** e **randomização**. [6]

impureza (*impurity*) Uma medida de incerteza dos nós de uma árvore de classificação. [11]

inclinação (*slope*) A mudança no valor esperado de uma variável resposta para uma dada mudança na variável preditora. O parâmetro β_1 em um modelo de regressão. [9]

indução (*induction*) O processo de raciocínio científico que progride do específico (este porco é amarelo) para o geral (todos os porcos são amarelos). *Comparar com* **dedução**. [4]

independente (*independent*) Dois eventos são ditos independentes se a probabilidade de um ocorrer não está relacionada de forma alguma à probabilidade de ocorrência do outro. [1, 6]

inferência (*inference*) Uma conclusão que, logicamente, surge de observações e premissas. [4]

inferência bayesiana (*Bayesian inference*) Umas das três grandes estruturas de análises estatísticas. Com frequência, é chamada de probabilidade inversa. A inferência bayesiana é usada para estimar a probabilidade de uma hipótese estatística, de acordo com os (ou condicionada pelos) dados disponíveis. O resultado de uma inferência bayesiana é uma nova distribuição de probabilidades (ou *a posteriori*). *Comparar com* as análises **frequentista, paramétrica** e de **Monte Carlo**. [5]

inferência frequentista (*frequentist inference*) Uma abordagem de inferência estatística na qual as probabilidades das observações são estimadas a partir de funções que assumem um número infinito de tentativas. *Comparar com* **inferência bayesiana**. [1, 3]

interação (*interaction*) Os efeitos conjuntos de dois ou mais fatores experimentais. Especificamente, a interação representa uma resposta que não pode ser predita apenas conhecendo o efeito principal de cada fator, de maneira isolada. As interações resultam em efeitos não aditivos. Ela é um importante componente dos experimentos multifatoriais. [7, 10]

intercepto (*intercept*) O ponto em que a linha de regressão cruza o eixo-*x*; ele é o valor esperado de *Y* quando *X* é igual a zero. [9]

interpolação (*interpolation*) A estimativa de valores não observados dentro da gama de valores observados. Em geral, é mais confiável que a **extrapolação**. [9]

interseção (*intersection*) Na teoria de conjuntos, os elementos que são comuns a dois ou mais conjuntos. Na teoria de probabilidades, a de que dois eventos ocorram ao mesmo tempo. Em ambos os casos, a interseção é escrita com o símbolo $\cap$. *Comparar com* **união**. [1]

intervalo (*interval*) Um conjunto de números com um mínimo e um máximo definidos. [2]

intervalo aberto (*open interval*) Um intervalo que não inclui seus extremos. Desta forma, o intervalo (0, 1) é o conjunto de todos os números entre 0 e 1, com exceção de 0 e 1. *Comparar com* **intervalo fechado**. [2]

intervalo de confiança (*confidence interval*) O intervalo que inclui a verdadeira média da população *n*% das vezes. O termo intervalo de confiança sempre deve ser classificado em *n*%, como em intervalo de confiança de 95%, intervalo de confiança de 50%, etc. Os intervalos de confiança são calculados usando estatística frequentista. Comumente são mal interpretados como representando uma chance de *n*% de que a média esteja dentro do intervalo de confiança. *Comparar com* **intervalo de credibilidade**. [3]

intervalo de credibilidade (*credibility interval*) Um intervalo que representa *n*% de chances de que a média da população esteja neste intervalo. Como no intervalo de confiança, o de credibilidade sempre deve ser classificado: intervalo de credibilidade de 95%, intervalo de credibilidade de 50%, etc. Esses intervalos são determinados conforme a estatística bayesiana. *Comparar com* **intervalo de confiança**. [3, 5]

intervalo fechado (*closed interval*) Um intervalo que inclui seus pontos finais. Assim, o intervalo fechado [0, 1] é o conjunto de todos os números entre 0 e 1, incluindo 0 e 1. *Comparar com* **intervalo aberto**. [2]

intervalo de predição inversa (*inverse prediction interval*) Fundamentado em um modelo de regressão, é a amplitude de valores possíveis do *X* (ou preditor) para um dado valor do Y (ou resposta). [9]

intervalo de predição simultânea (*simultaneous prediction interval*) Um ajuste em um intervalo de confiança de análise de regressão que é necessário caso seja feita uma estimativa de múltiplos valores. [9]

---

jackknife (*jackknife*) Uma ferramenta para vários propósitos. Na estatística, o jackknife é um processo de aleatorização (ou Monte Carlo) no qual uma observação é excluída do conjunto de dados, e a estatística é então recalculada. Isto é repetido tantas vezes quanto há de observações. A estatística teste, calculada a partir dos dados originais, é comparada com a distribuição da estatística teste obtida pela repetida aplicação do jackknife, fornecendo um valor de *P* exato. *Ver também* ***bootstrap*** e **função de influência**. [5, 12]

Lei dos Grande Números (*Law of Large Numbers*) Um teorema fundamental da probabilidade, a Lei dos Grandes Números diz que quanto maior a amostra, mais próxima a média aritmética se torna à expectativa probabilística de uma dada variável. [3]

Leptocúrtica (*leptokurtic*) Uma distribuição de probabilidades é dita leptocúrtica se ela tiver relativamente poucos valores no centro da distribuição e caudas pouco pesadas. A curtose de uma distribuição leptocúrtica é maior que 0. *Comparar com* **platicúrtica**. [3]

MANOVA *Ver* **análise de variância multivariada**. [12]

matriz (*matrix*) Um objeto matemático (frequentemente usado como variável) constituído por muitas observações (linhas) e muitas variáveis (colunas). [12]

matriz de classificação (*dassification matrix*) Um resultado de uma análise discriminante. Uma tabela onde as linhas representam os grupos observados a que cada observação pertence e as colunas representam a predição do grupo a que cada observação pertence. Os valores nesta tabela representam o número de observações colocadas em cada par (observado, predito). Em uma classificação perfeita (uma análise discriminante sem nenhum erro), somente os elementos da diagonal da matriz de classificação terão valores diferentes de zero. [12]

matriz de colunas ortonormais (*column-orthonormal matrix*) Uma matriz $m \times n$ cujas colunas foram normalizadas de maneira que a distância euclidiana de cada coluna = 1. [A]

matriz de gráfico de dispersão (*scatterplot matrix*) Um gráfico feito com vários gráficos de dispersão é usado para ilustrar dados multivariados. Cada gráfico é de dispersão simples e eles são agrupados por linhas (com variáveis *Y* em comum) e por colunas (com variáveis X em comum). O arranjo desses gráficos em uma matriz ilustra todas as relações bivariadas possíveis entre todas as variáveis medidas. [8]

matriz de variância e covariância (*variance-covariance matrix*) Uma matriz quadrada C (n linhas $\times$ n colunas, onde n é o número de medidas ou observações individuais para cada observação multivariada) em que os elementos da diagonal $c_{i,i}$ são as variâncias de cada variável e os elementos fora da diagonal $c_{i,j}$ são as covariâncias entre a variável i e a j. [9, 12]

matriz diagonal (*diagonal matrix*) Uma matriz quadrada em que todos os valores, exceto os da diagonal principal, são iguais a zero. [12]

matriz identidade (*identity matrix*) Uma matriz diagonal em que todos os termos da diagonal são = 1 e todos os outros = 0. [A]

matriz inversa (*inverse matrix*) Uma matriz $\mathbf{A}^{-1}$ tal que $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, a matriz identidade. [A]

matriz quadrada (*squared matrix*) Uma matriz com o mesmo número de linhas e de colunas. [12]

matriz triangular (*triangular matrix*) Uma matriz na qual todos os valores em um dos lados da diagonal são = 0. [A]

matriz triangular inferior (*lower triangular matrix*) Uma matriz na qual todos os elementos acima da diagonal são = 0. [A]

matriz triangular superior (*upper triangular matrix*) Uma matriz na qual todos os elementos abaixo da diagonal = 0. [A]

máxima verossimilhança (*maximum likelihood*) O valor mais provável da distribuição de verossimilhanças é calculado pegando a derivada da verossimilhança e determinando quando a derivada é igual a 0. Para muitos métodos estatísticos paramétricos ou frequentistas, a estatística teste assintótica é o valor de máxima verossimilhança daquela estatística. [5]

média (*avearage – mean*) Em inglês, existem duas palavras que podem ser traduzidas como média, *average* e *mean*. Portanto, a média pode ter 2 significados: 1) no contexto de *average*, indica a média aritmética de uma variável. Com frequência é usada de maneira incorreta para indicar o valor mais comum de uma variável. *Comparar com* **moda, mediana** e com o segundo significado de **média.** 2) no contexto de *mean*, é o valor mais provável de uma variável aleatória ou de um conjunto de observações. A média deve ser classificada em: aritmética, geométrica ou harmônica. Se não for classificada, assume-se que é a média aritmética. *Comparar com* **moda** e **mediana.** [3]

média ajustada (*adjusted mean*) O valor esperado de um grupo tratamento caso todos os grupos tenham a mesma covariável média. [10]

média aritmética (*arithmetic mean*) A média de um conjunto de observações de uma dada variável. Ela é calculada como a soma de todas as observações dividida pelo número de observações, e é simbolizada por uma barra horizontal sobre o nome da variável, por exemplo, $\overline{Y}$. [3]

média geométrica (*geometric mean*) A n-ésima raiz do produto de n observações. Uma média útil da expectativa para variáveis log-normal. A média geométrica é sempre menor que ou igual à média aritmética. [3]

média harmônica (*harmonic mean*) É o inverso da média aritmética dos inversos das observações. A média harmônica é uma estatística útil para calcular o tamanho efetivo da população (o tamanho de uma população com acasalamentos completamente aleatórios). A média harmônica sempre é menor que ou igual à média geométrica. [3]

mediana (*median*) O valor de um conjunto de observações que está exatamente no centro de todos os valores: metade dos valores são menores e metade são maiores. O cento do quinto decil, ou o 50-ésimo percentil. *Comparar com* **média** e **moda.** [3]

metadados (*metadata*) Dados sobre dados. A prosa descritiva que descreve todas as variáveis e características de um conjunto de dados. [8]

método de Bonferroni (*Bonferroni method*) Um ajuste do valor crítico, α, ao qual a hipótese nula será rejeitada. Ele é usado quando são feitas observações múltiplas, e é calculado como o α (geralmente igual a 0,05) dividido pelo número de comparações. *Comparar com* o **método de Dunn-Sidak.** [10]

método de Dunn-Sidak (*Dunn-Sidak method*) Um ajuste do valor crítico, α, no qual a hipótese nula será rejeitada. É usado quando múltiplas comparações são feitas e é calculado como $\alpha_{ajustado}$ é igual a $1-(1-\alpha_{nominal})^{1/k}$ onde $\alpha_{nominal}$ é o valor crítico do α (geralmente 0,05) e k é o número de comparações. *Comparar com* o **método de Bonferroni.** [10]

método hipotético-dedutivo (*hypothetico-deductive method*) O método científico creditado a Karl Popper. Como na indução, o método hipotético-dedutivo se inicia com uma observação individual. A observação pode ser predita por muitas hipóteses, cada uma das quais fornece predições adicionais. As predições são testadas uma a uma e as hipóteses alternativas vão sendo eliminadas até que reste apenas uma. Os praticantes do método hipotético-dedutivo nunca confirmam uma hipótese; eles apenas as rejeitam, ou falham em rejeitá-las. Quando os livros-texto se referem "ao método científico", geralmente estão se referindo ao método hipotético dedutivo. *Comparar com* **indução, dedução** e **inferência bayesiana.** [4]

métodos de Monte Carlo (*Monte Carlo methods*) Métodos estatísticos que recaem sobre a aleatorização ou o remanejamento dos dados, com frequência usando métodos de *bootstrap* ou de jackknife. O resultado de uma análise de Monte Carlo é um valor exato de probabilidade para os dados, dada a hipótese nula estatística. [5]

métrica (*metric*) Um tipo de medida de distância (ou dissimilaridade) que satisfaz quatro propriedades: (1) a distância mínima é igual a 0 e a entre objetos idênticos é igual a 0; (2) a distância entre dois objetos não idênticos é positiva; (3) as distâncias são simétricas: a do objeto a para o b é igual à distância do objeto b para o objeto a; e (4) as medidas das distâncias satisfazem a desigualdade dos triângulos. *Comparar com* **semimétrica** e **não métrica.** [12]

mínimos quadrados aparados (*least-trimmed squares*) Um método de regressão para minimizar o efeito de dados discrepantes. Calcula a inclinação da regressão após remover alguma proporção (especificada pelo analista) das maiores somas dos quadrados dos resíduos. [9]

moda (*mode*) O valor mais comum em um conjunto de observações. Coloquialmente, a média (ou média da população) em geral é confundida com a moda. *Comparar com* **média** e **mediana**. [3]

modelo (*model*) Uma função matemática que relaciona um conjunto de variáveis a outro por meio de um conjunto de parâmetros. Todos os modelos são errados, mas alguns são úteis. [11]

modelo de regressão linear (*linear regression model*) Um modelo estatístico no qual a relação entre o preditor e a variável resposta pode ser ilustrada por uma linha reta. Um modelo é dito linear porque seus parâmetros são constantes e afetam a variável preditora apenas aditiva ou multiplicativamente. *Comparar com* **modelo de regressão não linear**. [9]

modelo de regressão não linear (*non-linear regression model*) Um modelo estatístico no qual a relação entre as variáveis preditora e resposta não é uma linha reta e que as variáveis preditoras afetam a variável resposta de forma não aditiva ou não multiplicativa. *Comparar com* **modelo de regressão linear**. [9]

modelo de série temporal (*time series model*) Um modelo estatístico em que o tempo é incluído de manera explícita, como uma variável preditora. Muitos modelos de séries temporais também são **modelos autorregressivos**. [6]

modelo fatorial (*factor model*) O resultado de uma análise fatorial após os fatores comuns de suas cargas correspondentes terem sido padronizadas e da retenção apenas dos fatores comuns úteis (ou significativos). [12]

modelo log-linear (*log-linear model*) Um modelo estatístico em que o logaritmo dos parâmetros é linear. Por exemplo, o modelo $Y = aX^b$ é um modelo log-linear, pois o logaritmo transforma os resultados em um modelo linear: $\log(Y) = \log(a) + b \times \log(X)$. [11]

modelo misto (*mixed model*) Um modelo de ANOVA que inclui tanto efeitos aleatórios quanto fixos. [10]

modelos hierárquicos (*hierarchical models*) Um grupo de modelos estatísticos em que os parâmetros dos modelos mais simples são um subconjunto de parâmetros de modelos mais complexos. [11]

momento central (*central moment*) A média (aritmética) das diferenças entre cada observação de uma variável e a média aritmética daquela variável, elevada a *r*-ésima potência. O termo de momento central precisa sempre ser classificado: p. ex., primeiro momento central, segundo momento central, etc. O primeiro momento central é igual a 0 e o segundo momento central é a variância. [3]

mudança de escala (*change of scale*) A transformação de uma variável aleatória normal realizada multiplicando-a por uma constante. *Ver também* **operação de deslocamento**. [2]

multicolinearidade (*multicollinearity*) As correlações indesejadas entre duas ou mais variáveis preditoras. *Ver também* **colinearidade**. [7, 9]

multinormal (*multinormal*) Abreviação para normal multivariada. [12]

---

não métrica (*non-metric*) Um tipo de medida de distância (ou dissimilaridade) que satisfaz duas propriedades; (1) A distância mínima é igual a 0 e a entre objetos idênticos é igual a 0; (2) as distâncias são simétricas – ou seja, a distância do objeto *a* para o *b* é a mesma do objeto *b* para o *a*. *Comparar com* **métrica** e **semimétrica**. [12]

nível (*level*) O valor individual de um dado tratamento ou fator experimental em uma análise de variância ou em um delineamento tabular. [7, 11]

NMDS *Ver* **escalonamento multidimensional não métrico**. [12]

---

ordenação (*ordination*) Métodos de redução de dados multivariados através do arranjo das observações ao longo de um número de eixos menor que o de variáveis. [12]

ordenada (*ordinate*) A vertical, ou eixo-*y*, de um gráfico. O oposto da **abscissa**. [7]

operação de deslocamento (*shift operation*) A transformação de uma variável aleatória normal através da adição de uma constante. *Ver também* **mudança de escala**. [2]

ortogonal (*orthogonal*) Em uma análise de variância (ANOVA) multifatorial, é a propriedade de que todas as combinações dos tratamentos sejam representadas. Em um delineamento de regressão múltipla, é a propriedade de que todos os valores de uma variável preditora sejam encontrados em combinação com cada valor de outra variável preditora. [7,9]

ortogonal 2 (*orthogonal*) Em álgebra de matrizes, dois vetores cujo produto = 0 são ditos ortogonais. [A]

ortonormal (*orthonormal*) Uma transformação aplicada a uma matriz de forma que as distâncias euclidianas das linhas, colunas ou ambas sejam igual a 1. [12]

---

paradigma (*paradigm*) Uma estrutura de pesquisa sob a qual existe uma concordância geral. Thomas

Kuhn usou o termo "paradigma" ao descrever ciência "ordinária", que é a atividade de cientistas engajados em ajustar as observações aos paradigmas. Quando muitas observações não se ajustam, é hora de um novo paradigma. Kuhn chamou as mudanças de paradigmas de revoluções científicas. [4]

paramétrico (*parametric*) O pressuposto de que quantidades estatísticas podem ser estimadas usando distribuições de probabilidades com constantes definidas e fixas. O requisito de que vários testes estatísticos conformem uma distribuição de probabilidades conhecida e definível. [3]

parâmetro (*parameter*) Uma constante em distribuições e equações estatísticas que precisam ser estimadas a partir dos dados. [3, 4]

parâmetros parciais da regressão (*partial regression parameters*) Os parâmetros em um modelo de regressão múltipla cujos valores refletem a contribuição de todos os outros parâmetros no modelo. [9]

PCA *Ver* **análise de componentes principais.** [12]

percentil (*percentile*) A centésima parte de algo. Em geral é classificado como quinto percentil (um vigésimo), décimo percentil (decil), etc. [3]

permutação (*permutation*) O arranjo de objetos discretos. Diferente de combinação, a permutação é qualquer rearranjo de objetos. Assim, para o conjunto {1,2,3}, há seis permutações possíveis: {1,2,3}, {1,3,2}, {2,1,3}, {2,3,1}, {3,2,1}, {3,1,2}. Contudo, tudo isso representa a mesma combinação dos três objetos 1, 2, 3. Existem $n!$ permutações de n objetos, mas apenas

$$\frac{n!}{X!(n-X)!}$$

combinações de n objetos tirados X por vez. O símbolo "!" é o operador matemático para **fatorial.** *Ver também* **coeficiente binomial, combinação.** [2]

planilha eletrônica (*spreadsheet*) Um arquivo de computador em que cada linha representa uma única observação e cada coluna, uma única variável medida ou observada. A forma padrão na qual os dados ecológicos ambientais são colocados em meios eletrônicos. *Ver também* **arquivo simples.** [8]

platicúrtica (*platikurtic*) Uma distribuição de probabilidades é dita platicúrtica se ela tiver relativamente muitos valores no centro da distribuição e caudas relativamente finas. A curtose de uma distribuição platicúrtica é maior que 0. *Comparar com* **leptocúrtica.** [3]

plot (*plot*) Uma ilustração de uma variável ou um conjunto de dados. Tecnicamente, *plots* são elementos de gráficos. [8]

poder (*power*) A probabilidade de corretamente rejeitar uma hipótese nula estatística falsa. É calculado como $1 - \beta$, onde β é a probabilidade de cometer um erro Tipo II. [4]

porcentagem (*percentage*) Literalmente, "de 100" (do Latim *per centum*), a frequência relativa de um conjunto de observações multiplicada por 100. As porcentagens podem ser calculadas a partir de uma tabela de contingência, mas a análise estatística desta tabela sempre é baseada nas contagens de frequência relativa. [11]

pós-multiplicação e pós-produto (*postmultiplication and postproduct*) Em álgebra de matrizes, é o produto **BA** de duas matrizes quadradas **A** e **B**. [A]

posição (*location*) O local em uma distribuição de probabilidades onde a maioria das observações pode ser encontrada. *Comparar com* **dispersão.** [3]

pré-multiplicação e pré-produto (*premultiplication and preproduct*) Em álgebra de matrizes, é o produto **AB** de duas matrizes quadradas **A** e **B**. [A]

precisão (*precision*) O nível de concordância entre um conjunto de medidas sobre o mesmo indivíduo. Também é usada na inferência bayesiana para indicar a quantidade igual a 1/variância. *Comparar com* **acurácia.** [2, 5]

primeiro eixo principal (*first principal axis*) A nova variável, ou eixo, que resulta de uma análise de componentes principais e explica mais variabilidade nos dados que qualquer outro eixo. Também é chamado de primeiro componente principal. [12]

probabilidade (*probability*) A proporção de resultados em particular obtidos a partir de um dado número de tentativas, ou a proporção esperada para um número infinito de tentativas. [1]

probabilidade *a posteriori* (*posterior probability*) A probabilidade de uma hipótese, conforme os resultados de um experimento e qualquer conhecimento prévio. O resultado do Teorema de Bayes. *Comparar com* **probabilidade *a priori*.** [1]

probabilidade *a priori* (*prior probability*) A probabilidade de uma hipótese antes de um experimento ser conduzido. Isso pode ser determinado pela experiência prévia, intuição, opinião de especialistas ou revisão da literatura. *Comparar com* **probabilidade *a posteriori*.** [1]

probabilidade combinada de Fisher (*Fisher's combined probability*) Método para determinar se os resultados de múltiplas comparações são significantes, sem recorrer ao ajuste de taxa de erro por experimento. Ela é calculada como a soma dos logaritmos de todos os valores de P multiplicada

por –2. Esse teste estatístico é distribuído como uma variável aleatória de chi-quadrado com graus de liberdade igual a 2 vezes o número de comparações. [10]

probabilidade condicional (*conditional probability*) A probabilidade de um evento compartilhado quando os eventos simples associados não são independentes. Se o evento B já aconteceu e a probabilidade de observar o evento A depende do resultado de B, então dizemos que a probabilidade de A é *condicional* a B. Representamos isso como $P(A \mid B)$, e a calculamos como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad [1]$$

probabilidade de cauda (*tail probability*) A probabilidade de obter não apenas os dados observados, mas também possíveis dados mais extremos, mas que não foram observados. Também chamada de valor de P (valor de probabilidade) e calculada usando a função de distribuição cumulativa. *Ver também* **valor de probabilidade.** [2, 5]

probabilidade exata (*exact probability*) A probabilidade de obter somente o resultado observado. *Comparar com* **probabilidade de cauda.** [2, 5]

processos de Markov (*Markov process*) Uma sequência de eventos cuja probabilidade de um evento ocorrer depende somente da ocorrência de outro evento imediatamente predecessor. [3]

produto escalar (*scalar product*) Em álgebra de matrizes, é o produto de um vetor linha $1 \times n$ e um vetor coluna $n \times 1$. [A]

proporção (*proportion*) Frequência relativa. O número de observações de uma determinada categoria, dividido pelo número total de observações. As proporções podem ser calculadas a partir de tabelas de contingência, mas a análise estatística dessa tabela é sempre baseada na frequência de contagem original. [11]

proporção de variância explicada (PVE) (*proportion of variance explained – PEV*) A quantidade de variabilidade nos dados explicada por cada fator da análise de variância. É equivalente ao coeficiente de determinação do modelo de regressão. [10]

PVE *Ver* **proporção de variância explicada.** [10]

quadrado latino (*latin square*) Um delineamento de blocos aleatorizados no qual n tratamentos são definidos em um quadrado $n \times n$ e que cada tratamento aparece exatamente uma vez em cada linha e em cada coluna do campo. [7]

quadrado médio (*mean square*) A média das diferenças entre o valor de cada observação e a média aritmética dos valores, elevadas ao quadrado. Também é chamado de variância, ou segundo momento central. [3]

quantil (*quantile*) A n-ésima parte de algo. Tipos específicos de quantis incluem os decis, percentis e quartis. [3]

quartil (*quartile*) Um quarto de algo. Os quartis, superior e inferior, de um conjunto de dados são os 25% do fundo e os 25% do topo dos dados. [3]

r^2 ajustado (*adjusted r^2*) Quão bem um modelo de regressão se ajusta aos dados, descontando o número de parâmetros no modelo. É usado para evitar o superajuste de um modelo simplesmente incluindo variáveis preditoras adicionais. *Ver também* **coeficiente de determinação.** [9]

randomização (*randomization*) A atribuição de tratamentos experimentais a populações com base em tabelas de números, ou algoritmos, randômicos. *Comparar com* **casual.** [6]

rastro de auditoria (*audit trail*) O conjunto de arquivos e documentos associados que descrevem mudanças em um conjunto original de dados. Ela descreve todas as mudanças feitas no arquivo de dados original que resultaram no conjunto de dados tratado usado nas análises. [8]

razão de chances *a posteriori* (*posterior odds ratio*) Em uma análise bayesiana, a razão da distribuição de probabilidades *a posteriori* de duas hipóteses alternativas. [9]

razão de chances *a priori* (*prior odds ratio*) A razão da distribuição de probabilidades *a priori* de duas hipóteses alternativas. [9]

razão de Poisson (*rate parameter*) A constante que é inserida na fórmula de uma variável aleatória de Poisson. Além disso, a média, ou valor esperado, e a variância de uma variável aleatória de Poisson. [2]

razão-F de Fisher (*Fisher's F-ratio*) A estatística-F usada na análise de variância. [5]

RDA *Ver* **análise de redundância.** [12]

Regra dos 10 (*Rule of 10*) Um princípio-guia diz que você deveria ter no mínimo 10 réplicas por categoria observacional ou grupo de tratamento. A regra baseia-se na experiência, e não em algum axioma ou teorema matemático. [6]

regressão (*regression*) O método de análise usado para explorar relações de causa e efeito entre duas variáveis, cuja variável preditora é contínua. *Comparar com* **correlação.** [7, 9]

regressão gradativa (*stepwise regression*) Qualquer tipo de procedimento de regressão que envolve avaliar múltiplos modelos com parâmetros compartilhados. O objetivo é decidir quais parâmetros são úteis ao modelo de regressão final. Os exemplos de métodos usados na regressão gradativa incluem **eliminação regressiva** e **seleção progressiva**. [9]

regressão logística (*logistic regression*) Um modelo estatístico para estimar variáveis resposta categóricas a partir de uma variável preditora contínua. [9]

regressão múltipla (*multiple regression*) Um modelo estatístico que prediz uma resposta a partir de mais de uma variável preditora. Pode ser linear ou não linear. [9]

regressão quantílica (*quantile regression*) Um modelo de regressão sobre um subconjunto (um quantil definido pelo analista) dos dados. [9]

regressão robusta (*robust regression*) Um modelo de regressão que é relativamente pouco sensível a dados discrepantes ou valores extremos. Exemplos incluem a regressão de mínimos quadrados aparados e os estimadores-M. [9]

reificação (*reification*) Conversão de um conceito abstrato em um objeto material. [8]

réplica (*replicate*) Uma observação ou tentativa individual. A maioria dos experimentos possui réplicas múltiplas de cada categoria ou grupo de tratamento. [1, 7]

replicação (*replication*) O estabelecimento de múltiplas parcelas, observações ou grupos dentro do mesmo experimento ou tratamento observacional. Idealmente, a replicação é obtida pela aleatorização. [6]

resíduo (*residual*) A diferença entre um valor observado e seu valor previsto por um modelo estatístico. [8, 9]

resíduos normalizados (*studentized residuals*) *Ver* **resíduos padronizados**. [9]

resíduos padronizados (*standardized residuals*) Um resíduo transformado em um escore-Z. [9]

resultado (*outcome*) Do inglês *outcome*, é o resultado, geralmente de uma observação ou tentativa. [1]

resultado discreto (*discrete outcome*) Um resultado ou observação que pode assumir valores independentes ou inteiros. [1]

rotação (*rotation*) Em uma análise fatorial, rotação é uma combinação linear de fatores comuns que são facilmente interpretados. *Ver* **rotação oblíqua**, **rotação ortogonal** e **rotação varimax**. [12]

rotação oblíqua (*oblique rotation*) Uma combinação linear dos fatores comuns em uma análise fatorial que resulta em fatores comuns que podem ser correlacionados uns com os outros. *Comparar com* **rotação ortogonal** e **rotação varimax**. [12]

rotação ortogonal (*orthogonal rotation*) A combinação linear dos novos fatores comuns de uma análise fatorial que resulta em fatores comuns que não têm correlação uns com os outros. *Comparar com* **rotação oblíqua** e **rotação varimax**. [12]

rotação varimax (*varimax rotation*) Combinação linear dos fatores comuns em uma análise fatorial que minimiza a variância total dos fatores comuns. *Comparar com* **rotação oblíqua** e **rotação ortogonal**. [12]

ruído branco (*white noise*) Uma distribuição de erro em que os erros são independentes e não correlacionados (*ver* **IID**). É chamado de ruído branco por analogia com a luz: como a luz branca é uma mistura de todos os comprimentos de onda, o ruído branco é uma mistura de todas as distribuições de erro. [6]

---

scree plot (*scree plot*) Um gráfico de diagnóstico usado para decidir quantos componentes, fatores ou eixos são necessários reter em uma ordenação. Ele tem formato de encosta de montanha (ou declive de pedreira); os componentes úteis são os mais altos na encosta e os sem utilidade são os escombros, embaixo. [3]

seleção progressiva (*forward selection*) Um método de regressão gradativa (*stepwise regression*) que começa com um modelo que possui apenas um dentre vários parâmetros possíveis. Os parâmetros adicionais são testados e inseridos sequencialmente um de cada vez. *Ver também* **regressão gradativa** e *comparar com* **eliminação regressiva**. [9]

semimétrica (*semi-metric*) Um tipo de medida de distância (ou similaridade) que satisfaz três propriedades: (1) a distância mínima é igual a 0 e a entre objetos idênticos é igual a 0; (2) a distância entre dois objetos não idênticos é positiva; (3) as distâncias são simétricas – ou seja, a do objeto *a* até um objeto *b* é a mesma que do objeto *b* até o *a*. *Comparar com* **métrica** e **não métrica**. [12]

silogismo (*syllogism*) Uma dedução, ou conclusão, lógica (p. ex., este é um porco) é derivada de duas premissas (p. ex., 1, todos os porcos são amarelos; 2, este objeto é amarelo). Os silogismos foram primeiramente descritos por Aristóteles. Embora sejam lógicos, eles podem levar a conclusões errôneas (p. ex., e se o objeto amarelo na verdade for uma banana?). [4]

simétrico (*symmetric*) Um objeto cujos lados são imagens-espelho uma da outra. Uma matriz é dita simétrica se um espelho for alinhado ao longo de

sua diagonal; os valores (linha, coluna) acima da diagonal devem ser iguais aos valores (coluna, linha) abaixo da diagonal. [2,12]

soma dos produtos cruzados (SPC) (*sum of cross products – SCP*) A soma do produto da diferença entre cada observação e seu valor esperado. A soma dos produtos cruzados dividida pelo tamanho amostral é a **covariância**. [9]

soma dos quadrados (SQ) (*sum of squares – SS*) A soma das diferenças ao quadrado entre o valor de cada observação e a média aritmética de todos os valores. Usada extensivamente em cálculos da ANOVA. [3]

soma dos quadrados dos resíduos (*residual sum of squares*) O que é deixado (não é explicado) por um modelo de regressão ou ANOVA. É calculado como a soma dos desvios quadrados de cada observação até a média de todas as observações em seus grupos de tratamento, então essas somas são calculadas sobre todos os grupos de tratamento. Também chamada de **variação do erro** ou **variação residual**. [9, 10]

soma dos quadrados e produtos cruzados (SQPC) (*sum of squares and cross products – SSCP*) Uma matriz quadrada com o número de linha e número de colunas igual ao de variáveis em um conjunto de dados multivariado. Os elementos da diagonal dessa matriz são a soma dos quadrados de cada variável e os elementos fora da diagonal são a soma dos produtos cruzados de cada par de variáveis. As matrizes SQPC são usadas para calcular estatísticas-teste na MANOVA. [12]

SQPC *Ver* **soma dos quadrados e dos produtos cruzados**. [12]

subconjunto (*subset*) Uma coleção de objetos que também é parte de um grupo, ou conjunto, maior de objetos. *Ver também*, **conjunto**. [1]

---

tabela de contingência (*contingency table*) Uma forma de organizar, mostrar e analisar dados para ilustrar relações entre duas ou mais variáveis categóricas. Os valores em uma tabela de contingência são as frequências (contagens) de observações em cada categoria. [11]

tamanho do efeito (*effect size*) A diferença esperada entre as médias de grupos que estão sujeitos a diferentes tratamentos. [6]

taxa de erro por experimento (*experiment-wide error rate*) O valor crítico verdadeiro do α para o ponto em que a hipótese nula será rejeitada, após a correção de comparações múltiplas. Pode ser obtido usando os métodos de Bonferroni ou de Dunn-Sidak, dentre outros. [10]

tentativa (*trial*) Uma réplica ou observação visual. Muitas tentativas formam um experimento estatístico. [1]

tentativa de Bernoulli (*Bernoulli trial*) Um experimento que tem apenas dois resultados possíveis, como presença/ausência; vivo/morto; cara/coroa. Uma tentativa de Bernoulli resulta em uma variável aleatória de Bernoulli. [2]

Teorema de Bayes (*Bayes' Theorem*) A fórmula para calcular a probabilidade de uma hipótese de interesse, conforme os dados observados e qualquer conhecimento prévio acerca da hipótese:

$$P(\mathrm{H}|Y) = \frac{f(Y|\mathrm{H})P(\mathrm{H})}{P(Y)} \qquad [1]$$

Teorema do Limite Central (*Central Limit Theorem*) O resultado matemático que permite o uso de estatística paramétrica sobre dados que não têm distribuição normal. O Teorema do Limite Central mostra que qualquer variável aleatória pode ser transformada em uma variável aleatória normal. [2]

teoria (*theory*) Na ciência, uma teoria é uma coleção de conhecimentos aceitos que foram construídos através de repetidas observações e testes estatísticos de hipóteses. As teorias formam as bases dos **paradigmas**. [6]

teste bicaudal (*two-tailed test*) O teste da hipótese alternativa estatística diz que a estatística do teste observada não é igual ao valor esperado sob a hipótese nula estatística. As tabelas estatísticas e os programas estatísticos geralmente proveem valores críticos e de *P* associados para os testes bicaudais. *Comparar com* **teste unicaudal**. [5]

teste de bondade do ajuste (*goodness-of-fit test*) Teste estatístico usado para determinar a bondade do ajuste. [11]

teste de randomização (*randomization test*) Testes estatísticos que recaem no remanejamento dos dados, usando *bootstrap*, jackknife ou outra estratégia de re-amostragem. O remanejamento simula os dados que seriam coletados caso a hipótese nula estatística fosse verdadeira. *Ver também* **métodos de Monte Carlo, *bootstrap* e jackknife**. [5]

teste de razão de verossimilhanças (*likelihood ratio test*) Um método para testar entre diferentes hipóteses com base na razão de suas relativas probabilidades. [11]

teste unicaudal (*one-tailed test*) Um teste da hipótese estatística alternativa de que os valores obser-

vados da estatística teste é maior ou menor (e não ambos) que o valor esperado sob a hipótese nula estatística. Como as tabelas de estatística normalmente dão o resultado para um teste bicaudal, o valor da probabilidade unicaudal pode ser encontrado dividindo a probabilidade bicaudal por dois. *Comparar com* **teste bicaudal.** [5]

tolerância (*tolerance*) Um critério para decidir quando incluir ou não variáveis ou parâmetros em um procedimento de regressão gradativa. A tolerância reduz a multicolinearidade entre variáveis candidatas. [9]

total da coluna (*column total*) A soma marginal dos valores de uma coluna em uma tabela de contingência. *Ver também* **total da linha; total geral.** [11]

total da linha (*row total*) A soma marginal dos valores de uma única linha de uma tabela de contingência. *Ver também* **total da coluna e total geral.** [11]

total marginal (*marginal total*) A soma das frequências das linhas ou colunas de uma tabela de contingência. [7, 11]

traço (*trace*) A soma dos elementos da diagonal principal de uma matriz quadrada. [A]

transformação angular (*angular transformation*) A função $Y^* =$ arco-seno $\sqrt{Y}$, onde o arco-seno(X) é a função que retorna o ângulo θ para o qual o seno de $\theta = X$. Também chamada de **transformação do arco-seno** ou de **transformação do arco-seno da raiz quadrada.** [8]

transformação Box-Cox (*Box-Cox transformation*) Uma família de transformações de potências definida pela fórmula $Y^* = (Y^\lambda - 1)/\lambda$ (para $\lambda \neq 0$) e $Y^* = \ln(Y)$ (para $\lambda = 0$). O valor do λ que resulta no melhor ajuste a uma distribuição normal é usado para a transformação final dos dados. Esse valor deve ser obtido por métodos de computação intensiva. [8]

transformação da raiz quadrada (*square-root transformation*) A função $Y^* = \sqrt{Y}$. É usada com mais frequência para dados de contagem, que são variáveis aleatórias de Poisson. [8]

transformação do arco-seno (*arcsine transformation*) *Ver* **transformação angular.** [8]

transformação do arco-seno da raiz quadrada (*arcsine square-root transformation*) *Ver* **transformação angular.** [8]

transformação inversa (*reciprocal transformation*) A função $Y^* = 1/Y$ em geral é usada para dados hiperbólicos. [8]

transformação logarítmica (*logarithmic transformation*) A função $Y^* = \log(Y)$, onde log é o logaritmo em qualquer base útil. Com frequência, é usada quando as médias e variâncias são correlacionadas positivamente. [8, 9]

transformação *logit* (*logit transformation*) Uma transformação algébrica para converter os resultados de uma regressão logística, que produz uma curva em forma de S, em uma linha reta. [9]

transformação reversa (*back-transformed*) A consequência de mudar uma variável ou uma estimativa estatística de volta às unidades de medida originais. A transformação reversa só é necessária se as variáveis tiverem sido transformadas antes das análises estatísticas. *Ver também* **transformação.** [3, 8]

transposição (*transpose*) O rearranjo de uma matriz tal que as linhas da matriz original são iguais às colunas da nova (matriz transposta) e as colunas da matriz original são iguais às linhas da matriz transposta. Simbolicamente, para uma matriz **A** com elementos a_{ij}, a matriz transposta $\mathbf{A}^T$ tem os elementos $a'_{ij} = a_{ji}$. A transposição de matrizes é indicada ou pelo sobrescrito T ou pela aspa ('). [12]

tratamento (*treatment*) Um conjunto de categorias de variáveis preditoras usadas em um delineamento de análise de variância. Os tratamentos são equivalentes a fatores e são constituídos de níveis. [7]

---

união (*union*) Na teoria de conjuntos, são todos os elementos únicos obtidos pela combinação de dois ou mais conjuntos. Na teoria de probabilidades, é a probabilidade de dois eventos independentes ocorrerem. Em ambos os casos, a união é escrita com o símbolo $\cup$. *Comparar com* **intercessão.** [1]

---

valor crítico (*critical value*) O valor que um teste estatístico precisa ter para que sua probabilidade seja estatisticamente significativa. Cada teste estatístico é relacionado a uma distribuição de probabilidades que possui um valor crítico para um dado tamanho amostral e graus de liberdade. [5]

valor de *P* (*P-value*) *Ver* **valor de probabilidade.** [4]

valor de probabilidade (valor de *P*) (*probability value/ P-value*) A probabilidade de rejeitar uma hipótese nula estatística verdadeira também é chamada de probabilidade de erro Tipo I. [4]

valor esperado (*expected value*) O valor mais provável de uma variável aleatória também é chamado de expectativa de uma variável aleatória. [2]

valores singulares (*singular values*) Os valores ao longo da diagonal de uma matriz quadrada **W** tal que uma matriz **A** $m \times n$ possa ser re-expressa como o produto **VWU**, onde **V** é uma matriz $m \times n$

com a mesma dimensão de **A**, **U** e **W** são matrizes quadradas (dimensões $n \times n$); e **V** e **U** são matrizes com colunas ortonormais. [A]

variação (*variation*) Incerteza; diferença. [1]

variação dentro de grupos (*variation within groups*) *Ver* **soma dos quadrados dos resíduos.** [10]

variação do erro (*error variation*) O que permanece (não explicado) por um modelo de regressão ou de ANOVA. Também chamado de variação residual ou de soma dos quadrados dos resíduos, ele reflete erros de medidas e variação devido a causas que não são especificadas no modelo estatístico. [9, 10]

variação entre grupos (*variation among groups*) Em uma análise de variância, o desvio quadrado das médias dos grupos de tratamentos da média geral, somados para todos os grupos de tratamentos. Também é chamada de soma dos quadrados entre grupos. [10]

variação residual (*residual variation*) O que é deixado (não é explicado) por um modelo de regressão ou de ANOVA. Também é chamada de **variação do erro** ou de **soma dos quadrados dos resíduos.** [9, 10]

variância (*variance*) Uma medida de quão longe os valores observados diferem do valor esperado. [2]

variância amostral (*sample variance*) Uma estimativa sem tendência da variância de uma amostragem. Igual à soma dos quadrados dividida pelo tamanho amostral – 1. [3]

variável aleatória (*random variable*) A função matemática que atribui um valor numérico a cada resultado experimental. [12]

variável aleatória binomial (*binomial random variable*) O resultado de um experimento que consiste de múltiplas tentativas de Bernoulli. [2]

variável aleatória contínua (*continuous random variable*) O resultado de um experimento que pode assumir qualquer valor quantitativo. *Comparar com* **variável aleatória discreta.** [2]

variável aleatória de Bernoulli (*Bernoulli random variable*) O resultado de um experimento em que há somente dois resultados possíveis, como a presença/ausência; vivo/morto; cara/coroa. *Ver também* **variável aleatória.** [2]

variável aleatória de Poisson (*Poisson random variable*) O resultado discreto de um dado experimento conduzido em uma área finita ou em uma quantidade de tempo finita. Diferente da variável aleatória binomial, a de Poisson pode assumir qualquer valor inteiro. [2]

variável aleatória discreta (*discrete random variable*) O resultado de um experimento que assume apenas valores inteiros. *Comparar com* **variável aleatória contínua.** [1]

variável aleatória gaussiana (*gaussian random variable*) *Ver* **variável aleatória normal.** [2]

variável aleatória log-normal (*log-normal random variable*) Uma variável aleatória cujo logaritmo natural é uma variável aleatória normal. [2]

variável aleatória multinomial (*multinomial random variable*) Uma variável aleatória que pode assumir múltiplos valores. Ela é a extensão multivariada de uma variável aleatória binomial. [11]

variável aleatória normal (*normal random variable*) Uma variável aleatória cuja função de distribuição de probabilidades é simétrica em torno da média e que pode ser descrita por dois parâmetros, sua média (μ) e sua variância (σ^2). Também chamada de variável aleatória gaussiana. [2]

variável aleatória normal multivariada (*multivariate normal random variable*) O análogo de uma variável aleatória normal (ou gaussiana) para dados multivariados. Sua distribuição de probabilidades é simétrica ao redor de seu vetor médio, e é caracterizada por um vetor médio n-dimensional e por uma matriz de variância-covariância $n \times n$. [12]

variável aleatória normal padrão (*standard normal random variable*) Uma função de distribuição de probabilidades que produz uma variável aleatória normal com média = 0 e variância = 1. Se uma variável X é uma normal padrão, ela é escrita como $X \sim N(0,1)$. Em geral indicada por Z. [2]

variável aleatória uniforme (*uniform random variable*) Uma variável aleatória em que qualquer resultado é igualmente provável. [2]

variável categórica (*categoric variable*) Características que são classificadas em dois ou mais grupos distintos. *Comparar com* **variável contínua.** [7, 11]

variável contínua (*continuous variable*) Características que são medidas usando números inteiros ou números reais. *Comparar com* **variável categórica.** [7]

variável dependente (*dependent variable*) Em uma declaração de causa e efeito, a variável resposta, ou o objeto afetado a partir do qual se deseja determinar a causa. *Comparar com* **variável independente** e **variável preditora.** [7]

variável independente (*independent variable*) Em uma declaração de causa e efeito, é a variável preditora, ou o objeto que é postulado como o causador

do efeito observado. *Comparar com* **variável dependente** e **variável resposta.** [7]

variável preditora (*predictor variable*) O fator hipotético de causa. *Ver também* **variável independente** e *comparar com* **variável resposta.** [7]

variável resposta (*response variable*) O efeito hipotético de um fator causal. *Ver também* **variável dependente** e *comparar com* **variável preditora.** [7]

verossimilhança (*likelihood*) Uma distribuição empírica que permite quantificar a preferência entre hipóteses. Ela é proporcional à probabilidade de um conjunto específico de dados, dada uma determinada hipótese estatística ou um conjunto de parâmetros: L(hipótese | dados observados) = cP(dados observados | hipótese), mas diferente de uma distribuição de probabilidades, a verossimilhança não é restringida ao intervalo entre 0,0 e 1,0. Ela é um dos dois termos no numerador do Teorema de Bayes. *Ver também* **máxima verossimilhança.** [5]

vetor coluna (*column vector*) Uma matriz com muitas linhas, mas apenas uma coluna. *Comparar com* **vetor linha.** [12]

vetor linha (*row vector*) Uma matriz com muitas colunas, mas apenas uma linha. *Comparar com* **vetor coluna.** [12]

vetor médio (*mean vector*) O vetor com valores mais prováveis das variáveis aleatórias multivariadas, ou dados multivariados. [12]

Literatura Citada

Abramsky, Z., M. L. Rosenzweig and A. Subach. 1997. Gerbils under threat of owl predation: Isoclines and isodars. *Oikos* 78: 81–90. [7]

Albert, J. H. 1996. Bayesian selection of log-linearmodels. *Canadian Journal of Statistics* 24: 327–347. [11]

Albert, J. H. 1997. Bayesian testing and estimation of association in a two-way contingency table. *Journal of the American Statistical Association* 92: 685–693. [11]

Allison, T. and D. V. Cicchetti. 1976. Sleep in mammals: Ecological and constitutional correlates. *Science* 194: 732–734. [8]

Allran, J. W. and W. H. Karasov. 2001. Effects of atrazine on embryos, larvae, and adults of anuran amphibians. *Environmental Toxicology and Chemistry* 20: 769–775. [7]

Anscombe, F. J. 1948. The transformation of Poisson, binomial, and negative binomial data. *Biometrika* 35: 246–254. [8]

Arnould, A. 1895. *Les croyances fondamentales du bouddhisme; avec préf. et commentaries explicatifs.* Société théosophique, Paris. [1]

Arnqvist, G. and D. Wooster. 1995. Meta-analysis: Synthesizing research findings in ecology and evolution. *Trends in Ecology and Evolution* 10: 236–240. [10]

Barker, S. F. 1989. *The elements of logic*, 5th ed. McGraw Hill, New York. [4]

Beltrami, E. 1999. *What is random? Chance and order in mathematics and life.* Springer-Verlag, New York. [1]

Bender, E. A., T. J. Case and M. E. Gilpin. 1984. Perturbation experiments in community ecology: Theory and practice. *Ecology* 65:1–13. [6]

Berger, J. O. and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. *American Scientist* 76: 159–165. [5]

Berger, J. O. and R. Wolpert. 1984. *The likelihood principle.* Institute of Mathematical Statistics, Hayward, California. [5]

Bernardo, J., W. J. Resetarits and A. E. Dunham. 1995. Criteria for testing character displacement. *Science* 268: 1065–1066. [7]

Björkman, O. 1981. Responses to different quantum flux densities. Pp. 57–105 in O. L. Lange, P. S. Nobel, C. B. Osmond and H. Ziegler (eds.). *Encyclopedia of plant physiology*, new series, vol. 12A. Springer-Verlag, Berlin. [4]

Blume, J. D. and R. M. Royall. 2003. Illustrating the Law of Large Numbers (and confidence intervals). *American Statistician* 57: 51–57. [3]

Boecklen, W. J. and N. J. Gotelli. 1984. Island biogeographic theory and conservation practice: Species–area or specicus-area relationships? *Biological Conservation* 29: 63–80. [8]

Boone, R. D., D. F. Grigal, P. Sollins, R. J. Ahrens and D. E. Armstrong. 1999. Soil sampling, preparation, archiving, and quality control. Pp. 3–28 in G. P. Robertson, D. C. Coleman, C. S. Bledsoe and P. Sollins (eds.). *Standard soil methods for long-term ecological research.* Oxford University Press, New York. [8]

Box, G. E. P. and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26: 211–243. [8]

Boyer, C. B. 1968. *A history of mathematics.* John Wiley & Sons, New York. [2]

Breiman, L., J. H. Friedman, R. A. Olshen and C. G. Stone. 1984. *Classification and regression trees.* Wadsworth, Belmont, CA. [11]

Brett, M. T. and C. R. Goldman. 1997. Consumer versus resource control in freshwater pelagic food webs. *Science* 275: 384–386. [7]

Brezonik, P. L. and 12 others. 1986. Experimental acidification of Little Rock Lake, Wisconsin. *Water Air Soil Pollution* 31: 115–121. [7]

Brown, J. H. and G. A. Leiberman. 1973. Resource utilization and coexistence of seedeating desert rodents in sand dune habitats. *Ecology* 54: 788–797. [7]

Burnham, K. P. and D. R. Anderson. 2002. *Model selection and inference: a practical informationtheoretic approach*, 2nd ed. Springer-Verlag, New York. [5, 6, 7, 9]

Butler, M. A. and J. B. Losos. 2002. Multivariate sexual dimorphism, sexual selection, and adaptation in Greater Antillean *Anolis* lizards. *Ecological Monographs* 72: 541–559. [7]

Cade, B. S. and B. R. Noon. 2003. Agentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment* 1: 412–420. [9]

Cade, B. S., J. W. Terrell and R. L. Schroeder. 1999. Estimating effects of limiting factors with regression quantiles. *Ecology* 80: 311–323. [9]

Caffey, H. M. 1982. No effect of naturally occurring rock types on settlement or survival in the intertidal barnacle, *Tesseropora rosea* (Krauss). *Journal of Experimental Marine Biology and Ecology* 63: 119–132. [7]

Caffey, H. M. 1985. Spatial and temporal variation in settlement and recruitment of intertidal barnacles. *Ecological Monographs* 55: 313–332. [7]

Cahill, J. F., Jr., J. P. Castelli and B. B. Casper. 2000. Separate effects of human visitation and touch on plant growth and herbivory in an old-field community. *American Journal of Botany* 89: 1401–1409. [6]

Carlin, B. P. and T. A. Louis. 2000. *Bayes and empirical Bayes methods for data analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. [5]

Carpenter, S. R. 1989. Replication and treatment strength in whole-lake experiments. *Ecology* 70: 453–463. [6]

Carpenter, S. R., T. M. Frost, D. Heisey and T. K. Kratz. 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. *Ecology* 70: 1142–1152. [6, 7]

Carpenter, S. R., J. F. Kitchell, K. L. Cottingham, D. E. Schindler, D. L. Christensen, D. M. Post and N. Voichick. 1996. Chlorophyll variability, nutrient input, and grazing: Evidence from whole lake experiments. *Ecology* 77: 725–735. [7]

Caswell, H. 1988. Theory and models in ecology: a different perspective. *Ecological Modelling* 43: 33–44. [4, 6]

Clark, J. S., J. Mohan, J. Dietze and I. Ibanez. 2003. Coexistence: How to identify trophic trade-offs. *Ecology* 84: 17–31. [4]

Clarke, K. R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* 18: 117–143. [12]

Clarke, K. R. and R. H. Green. 1988. Statistical design and analysis for a "biological effects" study. *Marine Ecology Progress Series* 46: 213–226. [12]

Cleveland, W. S. 1985. *The elements of graphing data*. Hobart Press, Summit, NJ. [8]

Cleveland, W. S. 1993. *Visualizing data*. Hobart Press, Summit, NJ. [8]

Cochran, W.G. and G.M. Cox. 1957. *Experimental designs*, 2nd ed. John Wiley & Sons, New York. [7]

Cody, M.L. 1974. *Competition and the structure of bird communities*. Princeton University Press, Princeton, NJ. [6]

Colwell, R. K. and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B* 345: 101–118. [12]

Congdon, P. 2002. *Bayesian statistical modeling*. John Wiley & Sons, Chichester, UK. [9]

Connor, E. F. and E. D. McCoy. 1979. The statistics and biology of the species–area relationship. *American Naturalist* 113: 791–833. [8]

Connor, E. F. and D. Simberloff. 1978. Species number and compositional similarity of the Galapágos flora and avifauna. *Ecological Monographs* 48: 219–248. [12]

Craine, S. J. 2002. *Rhexia mariana* L. (Maryland Meadow Beauty) New England Plant Conservation Program Conservation and Research Plan for New England. New England Wild Flower Society, Framingham, MA. http://www.newfs.org/ [2]

Creel, S., J. E. Fox, A. Hardy, J. Sands, B. Garrott and R. O. Peterson. 2002. Snowmobile activity and glucocorticoid stress responses in wolves and elk. *Conservation Biology* 16: 809–814. [4]

Crisp, D. J. 1979. Dispersal and re-aggregation in sessile marine invertebrates, particularly barnacles. *Systematics Association* 11: 319–327. [3]

Darwin, C. 1875. *Insectivorous plants.* Appleton, New York. [1]

Davies, R. L. 1993. Aspects of robust linear regression. *Annals of Statistics* 21: 1843–1899. [9]

Day, R. W. and G. P. Quinn. 1989. Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs* 59: 433–463. [10]

De'ath, G. and K. E. Fabricius. 2000. Classification and regression trees: Apowerful yet simple technique for ecological data analysis. *Ecology* 81: 3178–3192. [11]

Denison, D. G. T., C. C. Holmes, B. K. Mallick and A. F. M. Smith. 2002. *Bayesian methods for nonlinear classification and regression.* John Wiley & Sons, Ltd, Chichester, UK. [11]

Dennis, B. 1996. Discussion: Should ecologists become Bayesians? *Ecological Applications* 6: 1095–1103. [4]

Diamond, J. 1986. Overview: Laboratory experiments, field experiments, and natural experiments. Pp. 3–22 in J. Diamond and T. J. Case (eds.). *Community ecology.* Harper & Row, Inc., New York. [6]

Doornik, J. A. and H. Hansen. 1994. An omnibus test for univariate and multivariate normality. Working paper, Nuffield College, Oxford University. http://www.nuff.ac.uk/Users/Doornik/papers/normal2.pdf [12]

Dunne, J.A., J. Harte and Kevin J. Taylor. 2003. Subalpine meadow flowering phenology responses to climate change: integrating experimental and gradient methods. *Ecological Monographs* 73: 69–86. [10]

Edwards, A. W. F. 1992. *Likelihood: Expanded edition.* The Johns Hopkins University Press, Baltimore. [5]

Edwards, D. 2000. Data quality assurance. Pp. 70–91 in W. K. Michener and J. W. Brunt (eds.). *Ecological data: Design, management and processing.* Blackwell Science Ltd., Oxford. [8]

Efron, B. 1982. The jackknife, the bootstrap, and other resampling plans. *Monographs of the Society of Industrial and Applied Mathematics* 38: 1–92. [5]

Efron, B. 1986. Why isn't everyone a Bayesian (with discussion). *American Statistician* 40: 1–11. [5]

Ellison, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6: 1036–1046. [3, 4]

Ellison, A. M. 2001. Exploratory data analysis and graphic display. Pp. 37–62 in S. M. Scheiner and J. Gurevitch (eds.). *Design and analysis of ecological experiments*, 2nd ed. Oxford University Press, New York. [8]

Ellison, A. M. 2004. Bayesian inference for ecologists. *Ecology Letters* (in press). [3, 4]

Ellison, A. M. and N. J. Gotelli. 2001. Evolutionary ecology of carnivorous plants. *Trends in Ecology and Evolution* 16: 623–629. [1]

Ellison, A. M., E. J. Farnsworth and R. R. Twilley. 1996. Facultative mutualism between red mangroves and root-fouling sponges in Belizean mangal. *Ecology* 77: 2431–2444 [10]

Ellison, A. M., N. J. Gotelli, J. S. Brewer, D. L. Cochran-Stafira, J. Kneitel, T. E. Miller, A. C. Worley and R. Zamora. 2003. The evolutionary ecology of carnivorous plants. *Advances in Ecological Research* 33: 1–74. [1]

Emerson, B. C., K. M. Ibrahim and G. M. Hewitt. 2001. Selection of evolutionary models for phylogenetic hypothesis testing using parametric methods. *Journal of Evolutionary Biology* 14: 620–631. [12]

Englund, G. and S.D. Cooper. 2003. Scale effects and extrapolation in ecological experiments. *Advances in Ecological Research* 33: 161–213. [6]

Farnsworth, E. J. 2004. Patterns of plant invasion at sites with rare plant species throughout New England. *Rhodora* (in press). [11]

Farnsworth, E. J. and A. M. Ellison. 1996a. Scaledependent spatial and temporal variability in biogeography of mangrove-root epibiont communities. *Ecological Monographs* 66: 45–66. [6]

Farnsworth, E.J. and A. M. Ellison. 1996b. Sun–shade adaptability of the red mangrove, *Rhizophora mangle* (Rhizophoraceae): Changes through ontogeny at several levels of biological organization. *American Journal of Botany* 83: 1131–1143. [4]

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783–791. [12]

FGDC (Federal Geographic Data Committee). 1997. FGDC-STD-005: Vegetation Classification Standard. U.S. Geological Survey, Reston, VA. [8]

FGDC (Federal Geographic Data Committee). 1998. FGDC-STD-001–1998: Content Standard for

Digital Geospatial Metadata (revised June 1998). U.S. Geological Survey, Reston, VA. [8]

Fienberg, S. E. 1970. The analysis of multidimensional contingency tables. *Ecology* 51:419–433. [11]

Fienberg, S. E. 1980. *The analysis of cross-classified categorical data*, 2nd ed. MIT Press, Cambridge, MA. [11]

Fisher, N .I. 1993. *Statistical analysis of circular data*. Cambridge University Press, Cambridge. [10]

Fisher, R. A. 1925. *Statistical methods for research workers*. Oliver & Boyd, Edinburgh. [5]

Flecker. A. S. 1996. Ecosystem engineering by a dominant detritivore in a diverse tropical stream. *Ecology* 77: 1845–1854. [7]

Foster, D. R., D. Knight, and J. Franklin. 1998. Landscape patterns and legacies resulting from large infrequent forest disturbance. *Ecosystems* 1: 497–510. [8]

Fox, D. R. 2001. Environmental power analysis: Anew perspective. *Environmetrics* 12: 437–448. [7]

Franck, D. H. 1976. Comparative morphology and early leaf histogenesis of adult and juvenile leaves of *Darlingtonia californica* and their bearing on the concept of heterophylly. *Botanical Gazette* 137: 20–34. [8]

Fretwell, S. D. and H. L. Lucas, Jr. 1970. On territorial behavior and other factors influencing habitat distribution in birds. *Acta Biotheoretica* 19: 16–36. [4]

Friendly, M. 1994. Mosaic displays for multiway contingency tables. *Journal of the American Statistical Association* 89: 190–200. [11]

Frost, T. M., D. L. De Angelis, T. F. H. Allen, S. M. Bartell, D. J. Hall and S. H. Hurlbert. 1988. Scale in the design and interpretation of aquatic community research. Pp. 229–258 in S. R. Carpenter (ed.). *Complex interactions in lake communities*. Springer-Verlag, New York. [7]

Gaines, S. D. and M. W. Denny. 1993. The largest, smallest, highest, lowest, longest, and shortest: Extremes in ecology. *Ecology* 74: 1677–1692. [8]

Garvey, J. E., E. A. Marschall and R. A. Wright. 1998. From star charts to stoneflies: Detecting relationships in continuous bivariate data. *Ecology* 79: 442–447. [9]

Gauch, H. G., Jr. 1982. *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge. [12]

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin. 1995. *Bayesian data analysis*. Chapman & Hall, New York. [5, 11]

Gifi, A. 1990. *Nonlinear multivariate analysis*. John Wiley & Sons, Chichester, UK. [12]

Gilbreth, F. B., Jr. and E. G. Carey. 1949. *Cheaper by the dozen*. Thomas Crowell. New York. [6]

Gill, J. A., K. Norris, P. M. Potts, T. G. Gunnarsson, P. W. Atkinson and W. J. Sutherland. 2001. The buffer effect and large-scale population regulation in migratory birds. *Nature* 412: 436–438. [4]

Gnanadesikan, R. 1997. *Methods for statistical data analysis of multivariate observations*, 2nd ed. John Wiley and Sons, London. [12]

Goldberg, D. E. and S. M. Scheiner. 2001. ANOVA and ANCOVA: Field competition experiments. Pp. 77–98 in S. Scheiner and J. Gurevitch (eds.). *Design and analysis of ecological experiments*, 2nd ed. Oxford University Press, New York. [7]

Goodman, D. 1975. The theory of diversity-stability relationships in ecology. *Quarterly Review of Biology* 50: 237–266. [11]

Gotelli, N. J. 2001. *A primer of ecology*, 3rd ed. Sinauer Associates, Sunderland, MA. [3]

Gotelli, N. J. and A. E. Arnett. 2000. Biogeographic effects of red fire ant invasion. *Ecology Letters* 3: 257–261. [6]

Gotelli, N. J. and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4: 379–391. [12]

Gotelli, N. J. and A. M. Ellison. 2002a. Biogeography at a regional scale: determinants of ant species density in New England bogs and forest. *Ecology* 83: 1604–1609. [6, 9, 12]

Gotelli, N. J. and A. M. Ellison. 2002b. Assembly rules for New England ant assemblages. *Oikos* 99: 591–599 [6, 9, 12]

Gotelli, N. J. and G. L. Entsminger. 2003. *EcoSim: Null models software for ecology*. Version 7. Acquired Intelligence Inc. & Kesey-Bear. Burlington, VT. Available for free at http://homepages.together.net/~gentsmin/ecosim.htm. [5]

Gotelli, N. J. and G. R. Graves. 1996. *Null models in ecology*. Smithsonian Institution Press, Washington, DC. [5, 11, 12]

Gould S. J. 1977. *Ontogeny and phylogeny*. Harvard University Press, Cambridge, MA. [8]

Gould, S. J. 1981. *The mismeasure of man.* W. W. Norton & Company, New York. [12]

Graham, M. H. 2003. Confronting multicollinearity in ecological multiple regression. *Ecology* 84: 2809–2815. [7, 9]

Green, M. D., M. G. P. van Veller and D. R. Brooks. 2002. Assessing modes of speciation: Range asymmetry and biogeographical congruence. *Cladistics* 18: 112–124. [8]

Gurevitch, J. and S. T. Chester, Jr. 1986. Analysis of repeated measures experiments. *Ecology* 67: 251–255. [10]

Gurevitch, J., P. S. Curtis and M. H. Jones. 2001. Meta-analysis in ecology. *Advances in Ecological Research* 3232: 199–247. [10]

Gurland, J. and R.C. Tripathi. 1971. Asimple approximation for unbiased estimation of the standard deviation. *American Statistician* 25: 30–32. [3]

Halley, J. M. 1996. Ecology, evolution, and 1/*f*noise. *Trends in Ecology and Evolution* 11: 33–37. [6]

Hamaker, J. E. 2002. Aprobabilistic analysis of the "unfair" Euro coin. http://www.isip.msstate.edu/publications/seminars/msstate_misc/2002/euro_coin/presentation_v0.pdf [11]

Hand, D. J. and C. C. Taylor. 1987. *Multivariate analysis of variance and repeated measures.* Chapman & Hall, London. [12]

Harris, R. J. 1985. *A primer of multivariate statistics.* Academic Press, New York. [12]

Harte, J., A. Kinzig and J. Green. 1999. Self-similarity in the distribution and abundance of species. *Science* 284: 334–336. [8]

Hartigan, J. A. 1975. *Clustering algorithms.* John Wiley & Sons, New York, New York. [12]

Harville, D. A. 1997. *Matrix algebra from a statistician's perspective.* Springer-Verlag, New York. [Appendix]

Hilborn, R. and M. Mangel. 1997. *The ecological detective: Confronting models with data.* Princeton University Press, Princeton, NJ. [4, 5, 6, 11]

Hill, M. O. 1973. Reciprocal averaging: An eigenvector method of ordination. *Journal of Ecology* 61: 237–249. [12]

Hoffmann-Jørgensen, J. 1994. *Probability with a view toward statistics.* Chapman & Hall, London. [2]

Holling, C. S. 1959. The components of predation as revealed by a study of small mammal predation of the European pine sawfly. *Canadian Entomologist* 91: 293–320. [4]

Horn, H.S. 1986. Notes on empirical ecology. *American Scientist* 74: 572–573. [4]

Horswell, R. 1990. *A Monte Carlo comparison of tests of multivariate normality based on multivariate skewness and kurtosis.* Ph.D. Dissertation, Louisiana State University, Baton Rouge, LA. [12]

Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics* 2: 360–378. [12]

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417–441, 498–520. [12]

Hotelling, H. 1936. Relations between two sets of variables. *Biometrika* 28: 321–377. [12]

Hubbard, R. and M. J. Bayarri. 2003. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *American Statistician* 57: 171–182. [4]

Huber, P. J. 1981. *Robust statistics.* John Wiley & Sons, New York. [9]

Huelsenbeck, J. P., B. Larget, R. E. Miller and F. Ronquist. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology* 51: 673–688. [12]

Hurlbert, S. H. 1971. The nonconcept of species diversity: Acritique and alternative parameters. *Ecology* 52: 577–585. [11]

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187–211. [6, 7]

Hurlbert, S. H. 1990. Spatial distribution of the montane unicorn. *Oikos* 58: 257–271. [3]

Inouye, B. D. 2001. Response surface experimental designs for investigating interspecific competition. *Ecology* 82: 2696–2706. [7]

Ives, A. R., B. Dennis, K. L. Cottingham and S. R. Carpenter. 2003 Estimating community stability and ecological interactions from time-series data. *Ecological Monographs* 73: 301–330. [6]

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences naturalles.* [12]

Jackson, D. A. 1993. Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches. *Ecology* 74: 2204–2214. [12]

Jackson, D. A. and K. M. Somers. 1991. Putting things in order: The ups and downs of detrended correspondence analysis. *American Naturalist* 137: 704–712. [12]

Jackson, D. A., K. M. Somers and H. H. Harvey. 1989. Similarity coefficients: measures of cooccurrence and association or simply measures of occurrence? *American Naturalist* 133: 436–453. [12]

Jaffe, M. J. 1980. Morphogenetic responses of plants to mechanical stimuli or stress. *Bio-Science* 30: 239–243. [6]

Jennions, M. D. and A. P. Møller. 2002. Publication bias in ecology and evolution: An empirical assessment using the "trim and fill" method. *Biological Reviews* 77: 211–222. [10]

Juliano, S. 2001. Non-linear curve fitting: Predation and functional response curves. Pp. 178–196 in S. M. Scheiner and J. Gurevitch (eds.). *Design and analysis of ecological experiments*, 2nd ed. Oxford University Press, New York. [9]

Kareiva, P. and M. Anderson. 1988. Spatial aspects of species interactions: the wedding of models and experiments. Pp. 38–54 in A. Hastings (ed.). *Community ecology*. Springer-Verlag, Berlin. [6]

Kass, R. E. and A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90: 773–795. [9]

Kingsolver, J. G. and D. W. Schemske. 1991. Path analyses of selection. *Trends in Ecology & Evolution* 6: 276–280. [9]

Knapp, R. A., K. R. Matthews and O. Sarnelle. 2001. Resistance and resilience of alpine lake fauna to fish introductions. *Ecological Monographs* 71: 401–421. [6]

Knüsel, L. 1998. On the accuracy of statistical distributions in Microsoft Excel 97. *Computational Statistics and Data Analysis* 26: 375–377. [8]

Koizol, J. A. 1986. Assessing multivariate normality: A compendium. *Communications in Statistics: Theory and Methods* 15: 2763–2783. [12]

Kramer, M. and J. Schmidhammer. 1992. The chi-squared statistic in ethology: Use and misuse. *Animal Behaviour* 44: 833–841. [7]

Kuhn, T. 1962. *The structure of scientific revolutions*. University of Chicago Press, Chicago. [4]

Lakatos, I. 1978. *The methodology of scientific research programmes*. 1978. Cambridge University Press, New York [4]

Lakatos, I. and A. Musgrave (eds.). 1970. *Criticism and the growth of knowledge*. Cambridge University Press, London. [4]

Lambers, H., F. S. Chapin III and T. L. Pons. 1998. *Plant physiological ecology*. Springer-Verlag, New York. [4]

Larson, S., R. Jameson, M. Etnier, M. Fleming and B. Bentzen. 2002. Low genetic diversity in sea otters (*Enhydra lutris*) associated with the fur trade of the 18th and 19th centuries. *Molecular Ecology* 11: 1899–1903. [3]

Laska, M. S. and J. T. Wootton. 1998. Theoretical concepts and empirical approaches to measuring interaction strength. *Ecology* 79: 461–476. [9]

Law, B. E., O. J. Sun, J. Campbell, S. Van Tuyl and P. E. Thornton. 2003. Changes in carbon storage and fluxes in a chronosequence of ponderosa pine. *Global Change Biology* 9: 510–524. [6]

Legendre, P. and M. J. Anderson. 1999. Distancebased redundancy analysis: Testing multi-species responses in multi-factorial ecological experiments. *Ecological Monographs* 69: 1–24. [12]

Legendre, P. and E. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271–280. [12]

Legendre, P. and L. Legendre. 1998. *Numerical ecology*. Second English edition. Elsevier Science BV, Amsterdam. [6, 11, 12, Appendix]

Leonard, T. 1975. Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society, Series B* (Methodological). 37: 23–37. [11]

Levings, S. C. and J. F. A. Traniello. 1981. Territoriality, nest dispersion, and community structure in ants. *Psyche* 88: 265–319. [3]

Levins, R. 1968. *Evolution in changing environments: Some theoretical explorations*. Princeton University Press, Princeton, NJ. [9]

Lichstein, J. W, T. R. Simons, S. A. Shriner, K. E. Franzreb. 2003. Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs* 72: 445–463. [6]

Lilliefors, H. W. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62: 399–402.

Lilliefors, H. W. 1969. On the Kolmogorov-Smirnov test for the exponential distribution with mean

unknown. *Journal of the American Statistical Association* 64: 387–389.

Lindley, D. V. 1964. The Bayesian analysis of contingency tables. *Annals of Mathematical Statistics* 35: 1622–1643. [11]

Loehle, C. 1987. Hypothesis testing in ecology: Psychological aspects and the importance of theory maturation. *Quarterly Review of Biology* 62: 397–409. [4]

Lomolino, M. V. and M. D. Weiser. 2001. Towards a more general species–area relationship: Diversity on all islands, great and small. *Journal of Biogeography* 28: 431–445. [8]

Looney, S. W. 1995. How to use tests for univariate normality to assess multivariate normality. *American Statistician* 49: 64–70. [12]

Losos., J. B. 1996. Phylogenetic perspectives on community ecology. *Ecology* 77: 1344–1354, 1996. [12]

MacArthur, R. H. 1962. Growth and regulation of animal populations. *Ecology* 43: 579. [4]

MacArthur, R. H. and E. O. Wilson. 1967. *The theory of island biogeography.* Princeton University Press, Princeton, NJ. [8]

MacKay, D. J. C. 2002. 140 heads in 250 tosses—suspicious? http://www.cs.toronto.edu/~mackay/euro.pdf [11]

MacNally, R. 2000a. Modelling confinement experiments in community ecology: Differential mobility among competitors. *Ecological Modelling* 129: 65–85. [6]

MacNally, R. 2000b. Regression and modelbuilding in conservation biology, biogeograph, and ecology: The distinction between—and reconciliation of—"predictive" and "explanatory" models. *Biodiversity and Conservation* 9: 655–671. [7]

Madigan, D. and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89: 1535–1546. [11]

Magurran, A. E. 2003. *Ecological diversity and its measurement.* Princeton University Press, Princeton, NJ. [11]

Magurran, A.E. and P.A. Henderson. 2003. Explaining the excess of rare species in natural species abundance distributions. *Nature* 422: 714–716. [2]

Manly, B. F. J. 1991. *Multivariate statistical methods: A primer.* Chapman & Hall, London. [12]

Mardia, K. 1970. Measures of multivariate skewness and kurtosis with applications *Biometrika* 78: 355–363. [12]

Margalef, R. 1958. Information theory in ecology. *General Systems* 3: 36–71. [11]

May, R. M. 1975. Patterns of species abundance and diversity. Pp. 81–120 in M. L. Cody and J. M. Diamond (eds.). *Ecology and evolution of communities.* Belknap, Cambridge, MA. [2]

Mayr, E. 1963. *Animal species and evolution.* Harvard University Press, Cambridge, MA. [8]

McArdle, B. H. and M. J. Anderson. 2001. Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology* 82: 290–297. [12]

McArdle, B. H., K. J. Gaston and J. H. Lawton. 1990. Variation in the size of animal populations: Patterns, problems, and artifacts. *Journal of Animal Ecology* 59: 439–354. [8]

McCullagh, P. and J. A. Nelder. 1989. *Generalized linear models*, 2nd ed. Chapman & Hall, London. [9, 10]

McCullough, B. D. and B. Wilson. 1999. On the accuracy of statistical procedures in Microsoft Excel 97. *Computational Statistics and Data Analysis* 31: 27–37. [8]

Mead, R. 1988. *The design of experiments: Statistical principles for practical applications.* Cambridge University Press, Cambridge. [7]

Mecklin, C. J and D. J. Mundfrom. 2003. On using asymptotic critical values in testing for multivariate normality. http://interstat.statjournals.net/interstat/articles/2003/articles/jo3001.pdf [12]

Merkt, R. E. and A. M. Ellison. 1998. Geographic and habitat-specific morphological variation of *Littoraria* (*Littorinopsis*) *angulifera* (Lamarck, 1822). *Malacologia* 40: 279–295. [12]

Michener, W. K and K. Haddad. 1992. Database administration. Pp. 4–14 in G. Lauff and J. Gorentz (eds.). *Data management at biological field stations and coastal marine laboratories.* Michigan State University Press, East Lansing, MI. [8]

Michener, W. K. 2000. Metadata. Pp. 92–116 in W. K. Michener and J. W. Brunt (eds.). *Ecological data: Design, management and processing.* Blackwell Science Ltd., Oxford. [8]

Michener, W. K., J. W. Brunt, J. Helly, T. B. Kirchner and S. G. Stafford. 1997. Non-geospatial metadata

for the ecological sciences. *Ecological Applications* 7: 330–342. [8]

Miller, J. A. 2003. Assessing progress in systematics with continuous jackknifing function analysis. *Systematic Biology* 52: 55–65. [12]

Mitchell, R. J. 1992. Testing evolutionary and ecological hypotheses using path-analysis and structural equation modeling. *Functional Ecology* 6: 123–129. [9]

Mooers, A. Ø., H. D. Rundle and M. C. Whitlock. 1999. The effects of selection and bottlenecks on male mating success in peripheral isolates. *American Naturalist* 153: 437–444. [8]

Moore, J. 1984. Parasites that change the behavior of their host. *Scientific American* 250: 108–115. [7]

Moore, J. 2001. *Parasites and the behavior of animals.* Oxford Series in Ecology and Evolution. Oxford University Press, New York. [7]

Murtaugh, P. A. 2002a Journal quality, effect size, and publication bias in meta-analysis. *Ecology* 83: 1162–1166. [4]

Murtaugh. P. A. 2002b On rejection rates of paired intervention analysis. *Ecology* 83: 1752–1761. [6, 7]

Newell, S. J. and A. J. Nastase. 1998. Efficiency of insect capture by *Sarracenia purpurea* (Sarraceniaceae), the northern pitcher plant. *American Journal of Botany* 85: 88–91. [1]

Niklas, K. J. 1994. *Plant allometry: The scaling of form and process.* University of Chicago Press, Chicago. [8]

Orlóci, L. 1978. *Multivariate analysis in vegetation research*, 2nd ed. Dr. W. Junk B.V., The Hague, The Netherlands. [12]

Osenberg, C. W., O. Sarnelle, S. D. Cooper and R. D. Holt. 1999. Resolving ecological questions through meta-analysis: Goals, metrics, and models. *Ecology* 80: 1105–1117. [10]

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Fifth Series* 50: 157–172. [11]

Pearson, K. 1901. On lines and planes of closest fit to a system of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series* 2: 557–572. [12]

Peres-Neto, P. R., D. A. Jackson and K. M. Somers. 2003. Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis. *Ecology* 84: 2347–2363. [12]

Petraitis, P. 1998. How can we compare the importance of ecological processes if we never ask, "compared to what?" Pp. 183–201 in W. J. Resetarits Jr. and J. Bernardo (eds.). *Experimental ecology: Issues and perspectives.* Oxford University Press, New York. [7, 10]

Petraitis, P. S., A. E. Dunham and P. H. Niewiarowski. 1996. Inferring multiple causality: The limitations of path analysis. *Functional Ecology* 10: 421–431. [9]

Pielou, E. C. 1981. The usefulness of ecological models: A stock-taking. *Quarterly Review of Biology* 56: 17–31. [4]

Pimm, S.L. and A. Redfearn. 1988. The variability of population densities. *Nature* 334: 613–614. [6]

Platt, J. R. 1964. Strong inference. *Science* 146: 347–353. [4]

Podani, J. and I. Miklós. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology*. 33: 3331–3343. [12]

Popper, K. R. 1935. *Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft.* J. Springer, Vienna. [4]

Popper, K. R. 1945. *The open society and its enemies.* G. Routledge & Sons, London. [4]

Potvin, C. 2001. ANOVA: Experimental layout and analysis. Pp. 63–76 in S. M. Scheiner and J. Gurevitch (eds). *Design and analysis of ecological experiments*, 2nd ed. Oxford University Press, New York. [6]

Potvin, C., M. J. Lechowicz and S. Tardif. 1990. The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures. *Ecology* 711: 1389–1400. [10]

Potvin, C., J. P. Simon and B. R. Strain. 1986. Effect of low temperature on the photosynthetic metabolism of the C^4 grass *Echinocloa crus-galli*. *Oecologia* 69: 499–506. [10]

Poulin, R. 2000. Manipulation of host behaviour by parasites: Aweakening paradigm? *Proceedings of the Royal Society Series B* 267: 787–792. [7]

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling. 1986. *Numerical recipes.*

Cambridge University Press, Cambridge. [9, Appendix]

Preston, F. W. 1948. The commonness and rarity of species. *Ecology* 29: 254–283. [2]

Preston, F. W. 1962. The canonical distribution of commonness and rarity: Part I. *Ecology*. 43: 185–215. [8, 9]

Preston, F. W. 1981. Pseudo-lognormal distributions. *Ecology* 62: 355–364. [2]

Price, M. V. and N. M. Waser. 1998. Effects of experimental warming on plant reproductive phenology in a subalpine meadow. *Ecology* 79: 1261–1271. [10]

Pynchon, T. 1973. *Gravity's rainbow*. Random House, New York. [2]

Quinn, G. and M. Keough. 2002. *Experimental design and data analysis for biologists*. Cambridge University Press, Cambridge. [7, 10]

Rao, A. R. and W. Tirtotjondro. 1996. Investigation of changes in characteristics of hydrological time series by Bayesian methods. *Stochastic Hydrology and Hydraulics* 10: 295–317. [7]

Rasmussen, P. W., D. M. Heisey, E. V. Nordheim and T. M. Frost. 1993. Time-series intervention analysis: Unreplicated large-scale experiments. Pp. 158–177 in S. Scheiner and J. Gurevitch (eds.). *Design and analysis of ecological experiments*. Chapman & Hall, New York. [7]

Real, L. 1977. The kinetics of functional response. *American Naturalist* 111: 289–300. [4]

Reckhow, K. H. 1996. Improved estimation of ecological effects using an empirical Bayes method. *Water Resources Bulletin* 32: 929–935. [7]

Reese, W.L. 1980. *Dictionary of philosophy and religion: Eastern and western thought*. Humanities Press, New Jersey. page 572 [4]

Rejmánek, M and R. Klinger. 2003. CANOCO 4.5 and some comparisons with PC-ORD and SYN-TAX. *Bulletin of the Ecological Society of America* 84: 69–74. http://www.esapubs.org/bulletin/backissues/084-2/084-2depts.pdf. [12]

Rice, J. and R. J. Belland. 1982. Asimulation study of moss floras using Jaccard's coefficient of similarity. *Journal of Biogecgraphy* 9: 411–419. [12]

Ripley, B. D. 1996. *Pattern recognition and neural networks*. Cambridge University Press, Cambridge. [11]

Rogers, D.J. 1972. Random search and insect population models. *Journal of Animal Ecology* 41: 369–383. [9]

Rohlf, F. J. and R. R. Sokal. 1995. *Statistical tables*, 3rd ed. W. H. Freeman & Company, New York. [3, 11]

Rosenzweig, M. L. and Z. Abramsky. 1997. Two gerbils of the Negev: Along-term investigation of optimal habitat selection and its consequences. *Evolutionary Ecology* 11: 733–756. [10]

Royston, J. P. Some techniques for assessing multivariate normality based on the Shapiro-Wilk *W. Applied Statistics* 32: 121–133. [12]

Sale, P. F. 1984. The structure of communities of fish on coral reefs and the merit of a hypothesis-testing, manipulative approach to ecology. Pp. 478–490 in D. R. Strong, Jr., D. Simberloff, L. G. Abele and A. B. Thistle (eds.). *Ecological communities: Conceptual issues and the evidence*. Princeton University Press, Princeton, NJ. [4]

Salisbury, F.B. 1963. *The flowering process*. Pergamon Press, Oxford. [6]

Sanderson, M. J. and A. C. Driskell. 2003. The challenge of constructing large phylogenetic trees. *Trends in Plant Science* 8: 374–379. [12]

Sanderson, M. J. and H. B. Shaffer. 2002. Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics* 33: 49–72. [12]

Scharf, F. S., F. Juanes and M. Sutherland. 1998. Inferring ecological relationships from the edges of scatter diagrams. *Ecology* 79: 448–460. [9]

Scheiner, S. M. 2001. MANOVA: Multiple response variables and multispecies interactions. Pp. 99–115 in S. M. Scheiner and J. Gurevitch (eds.). *Design and analysis of ecological experiments*, 2nd ed. Oxford University Press, Oxford. [12]

Schindler, D. W., K. H. Mills, D. F. Malley, D. L. Findlay, J. A. Shearer, I. J. Davies, M. A. Turner, G. A. Lindsey and D. R. Cruikshank. 1985. Long-term ecosystem stress: The effects of years of experimental acidification on a small lake. *Science* 228: 1395–1401. [7]

Schluter, D. 1990. Species-for-species matching. *American Naturalist* 136: 560–568. [5]

Schluter, D. 1995. Criteria for testing character displacement response. *Science* 268: 1066–1067. [7]

Schluter, D. 1996. Ecological causes of adaptive radiation. *American Naturalist* 148: S40–S64. [8]

Schoener, T.W. 1991. Extinction and the nature of the metapopulation: Acase system. *Acta Oecologia* 12: 53–75. [6]

Schroeder, R. L. and L. D. Vangilder. 1997. Tests of wildlife habitat models to evaluate oak mast production. *Wildlife Society Bulletin* 25: 639–646. [9]

Schroeter, S. C., J. D. Dixon, J. Kastendiek, J. R. Bence and R. O. Smith. 1993. Detecting the ecological effects of environmental impacts: Acase study of kelp forest invertebrates. *Ecological Applications* 3: 331–350. [7]

Shipley, B. 1997. Exploratory path analysis with applications in ecology and evolution. *American Naturalist* 149: 1113–1138. [9]

Shrader-Frechette, K. S. and E. D. McCoy. 1992. Statistics, costs and rationality in ecological inference. *Trends in Ecology and Evolution* 7: 96–99. [4]

Shurin, J. B. E. T. Borer, E. W. Seabloom, K. Anderson, C. A. Blanchette, B. Broitman, S. D. Cooper, B. S. Halpern. 2002. A cross-ecosystem comparison of the strength of trophic cascades. *Ecology Letters* 5: 785–791. [10]

Silvertown, J. 1987. *Introduction to plant ecology*, 2nd ed. Longman, Harlow, U.K. [7]

Simberloff, D. 1978. Entropy, information, and life: Biophysics in the novels of Thomas Pynchon. *Perspectives in Biology and Medicine* 21: 617–625. [2]

Simberloff, D. and L. G. Abele. 1984. Conservation and obfuscation: Subdivision of reserves. *Oikos* 42: 399–401. [8]

Sjögren-Gulve, P. and T. Eberhard. 2000. *The use of population viability analyses in conservation planning*. Munksgaard, Copenhagen. [6]

Smith, G. D. and S. Ebrahim. 2002. Data dredging, bias, or confounding. *British Medical Journal* 325: 1437–1438. [8]

Sneath, P. H. A. and R. R. Sokal. 1973. *Numerical taxonomy: The principles and practice of numerical classification*. W. H. Freeman & Company, San Francisco. [12]

Sokal, R. R. and F. J. Rohlf. 1995. *Biometry*, 3rd ed. W. H. Freeman & Company, New York. [3, 4, 5, 8, 9, 10, 11]

Somerfield, P. J., K. R. Clarke and F. Olsgard. 2002. Acomparison of the power of categorical and correlational tests applied tc community ecology data from gradient studies. *Journal of Animal Ecology* 71: 581–593. [12]

Sousa, W. P. 1979. Disturbance in marine intertidal boulder fields: the nonequilibrium maintenance of species diversity. *Ecology* 60: 1225–1239. [4]

Spearman, C. 1904. "General intelligence," objectively determined and measured. *American Journal of Psychology* 15: 201–293. [12]

Spiller, D. A. and T. W. Schoener. 1995. Longterm variation in the effect of lizards on spider density is linked to rainfall. *Oecologia* 103: 133–139. [6]

Spiller, D. A. and T. W. Schoener. 1998. Lizards reduce spider species richness by excluding rare species. *Ecology* 79: 503–516. [6]

Stewart-Oaten, A. and J. R. Bence. 2001. Temporal and spatial variation in environmental impact assessment. *Ecological Monographs* 71: 305–339. [6, 7]

Stewart-Oaten, A., J. R. Bence and C. W. Osenberg. 1992. Assessing effects of unreplicated perturbations: No simple solutions. *Ecology* 73: 1396–1404. [7]

Strauss, R. E. 1982. Statistical significance of species clusters in association analysis. *Ecology* 63: 634–639. [12]

Sugar, C. A. and G. M. James. 2003. Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association* 98: 750–763. [12]

Sugihara, G. 1980. Minimal community structure: an explanation of species abundance patterns. *American Naturalist* 116: 770–787. [2]

Thompson, J. D., G. Weiblen, B. A. Thomson, S. Alfaro and P. Legendre. 1996. Untangling multiple factors in spatial distributions: Lilies, gophers, and rocks. *Ecology* 77: 1698–1715. [9]

Trexler, J. C. and Travis, J. 1993. Nontraditional regression analysis. *Ecology* 74: 1629–1637. [9]

Tufte, E. R 1986. *The visual display of quantitative information*. Graphics Press, Cheshire, CT. [8]

Tufte, E. R. 1990. *Envisioning information*. Graphics Press, Cheshire, CT. [8]

Tukey, J. W. 1977. *Exploratory data analysis*. Addison-Wesley, Reading, MA. [8]

Turchin, P. 2003. *Complex population dynamics: A theoretical/empirical synthesis*. Princeton University Press, Princeton, NJ. [6]

Ugland, K. I. and J. S. Gray. 1982. Lognormal distributions and the concept of community equilibrium. *Oikos* 39: 171–178. [2]

Underwood, A. J. 1986. The analysis of competition by field experiments. Pp. 240–268 in J. Kikkawa and D. J. Anderson (eds.). *Community ecology: Pattern and process*. Blackwell, Melbourne. [7]

Underwood, A. J. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. *Ecological Applications* 4: 3–15. [6, 7]

Underwood, A. J. 1997. *Experiments in ecology: Their logical design and interpretation using analysis of variance*. Cambridge University Press, Cambridge. [7, 10]

Underwood, A. J. and P. S. Petraitis. 1993. Structure of intertidal assemblages in different locations: How can local processes be compared? Pp. 39–51 in R. E. Ricklefs and D. Schluter (eds.). *Species diversity in ecological communities: Historical and geographical perspectives*. University of Chicago Press, Chicago. [10]

Varis, O. and S. Kuikka. 1997. BeNe-EIA: A Bayesian approach to expert judgement elicitation with case studies on climate change impacts on surface waters. *Climatic Change* 37: 539–563. [7]

Venables, W. N. and B. D. Ripley. 2002. *Modern applied statistics with S*, 4th ed. Springer-Verlag, New York. [9, 11]

Wachsmuth, A., L. Wilkinson and G. E. Dallal. 2003. Galton's bend: Apreviously undiscovered nonlinearity in Galton's family stature regression data. *The American Statistician* 57: 190–192. [9]

Watson, J. D. and F. H. Crick. 1953. Molecular structure of nucleic acids. *Nature* 171: 737–738. [4]

Weiner, J. and O. T. Solbrig. 1984. The meaning and measurement of size hierarchies in plant populations. *Oecologia* 61: 334–336. [3]

Weisberg S. 1980. *Applied linear regression*. John Wiley & Sons, New York. [9]

Werner, E. E. 1998. *Ecological experiments and a research program in community ecology*. Pp. 3–26 in W. J. Resetarits Jr. and J. Bernardo (eds.). *Experimental ecology: Issues and perspectives*. Oxford University Press, New York. [7]

Whittaker, R. H. 1967. Vegetation of the Great Smoky Mountains. *Ecological Monographs* 26: 1–80. [12]

Wiens, J. A. 1989. Spatial scaling in ecology. *Functional Ecology* 3: 385–397. [6]

Williams, D. A. 1976. Improved likelihood ratio tests for complete contingency tables. *Biometrika* 39: 274–289. [11]

Williams, M. R. 1996. Species–area curves: The need to include zeroes. *Global Ecology and Biogeography Letters* 5: 91–93. [9]

Williamson, M., K. J. Gaston and W. M. Lonsdale. 2001. The species–area relationship does not have an asymptote! *Journal of Biogeography* 28: 827–830. [8]

Willis, E. O. 1984. Conservation, subdivision of reserves, and the anti-dismemberment hypothesis. *Oikos* 42: 396–398. [8]

Wilson, J. B. 1993. Would we recognise a brokenstick community if we found one? *Oikos* 67: 181–183. [2]

Winer, B. J., D. R. Brown and K. M. Michels. 1991. *Statistical principles in experimental design*, 3rd ed. McGraw-Hill, New York. [7, 10]

Wolda, H. 1981. Similarity indices, sample size, and diversity. *Oecologia* 50: 296–302 [12]

Índice

Números de página em *itálico* referem-se a informações em uma tabela ou ilustração. A letra n refere-se a uma informação de nota de rodapé, sendo seguida do número da nota no respectivo capítulo.

A

Abscissa (horizontal / eixo-*x*), 183, 183n1
Ácido sulfúrico, 213, 213n9
Agências de fomento, 225-226, 225-226n1
Agrupamento
  aglomerativo *vs.* divisivo, 448-450
  análise de, 447-448, *448-449*
  divisivo, 448-450
  hierárquico *vs.* não hierárquico, 449-451
  K-médias, 451
Álcool, 349-350n8
Aleatorização, ou Randomização, 115
Aletorização
  controle da, 213
  delineamentos experimentais e, 181-182
  em experimentos, 172-176, 172-173n9-10, 174-175n11, 175n12
  restrições sobre, 193-196, 195n4, 196
  testes de, 128-129n2, 131
Alfa (α). *Ver* erro Tipo I
Álgebra de matrizes, 402, 405-406, 465-475
  adição e subtração, 465-467
  autovalores, 471-475
  autovetores, 471-475
  multiplicação, 466-468
  transposição, 467-475
Amostragem, 23, *27*, 27-30, *28-29*, 28-29n6
  aleatória, 172-173n9
  casual, 172-173n9
  delineamento de, 181-182, 220-221
  esforço de, 200
  estudo, 181-182
Amostras, *27*, 27-30, *28-29*, 28-29n6, 312-314
  amplitude, 185-186
  análises de, 125-128, 126-127n1
  covariância, 263-264
  desvio-padrão das, 84-87, 84-85n5
  escala espacial das, 126-127
  população, 126-127
  tamanho/correção de valores para, 378-379
  variância, 84-85
Amplitude, 234-236, *235*
Análise
  agrupamento, 447-449
  amostras, 125-128, 126-127n1, 127-128
  análise de correspondência canônica (CCA), 455-456
  bayesiana. *Ver* análise bayesiana
  caminhos, 257, 297-301, 298-299, 298n16, 302-303
  componentes principais. *Ver* Análise de componentes principais
  coordenadas principais. *Ver* Análise de coordenadas principais
  correspondência. *Ver* análise de correspondência
  correspondência destendenciada (DCA), 443-444
  discriminante. *Ver* Análise discriminante
  espécies indicadores com dois fatores (TWINSPAN), 449-450
  fatorial. *Ver* Análise fatorial
  filogenética, 454-455
  gradiente indireto, 439
  impacto ambiental, 161-162n4
  intervenção randomizada (RIA), 213-215
  modelos mistos, 335
  Monte Carlo. *Ver* análise de Monte Carlo
  multivariada, 402
  paramétrica. *Ver* Análise paramétrica
  tabela de contingência, 218-219
Análise de Agrupamentos, 447-448, *448-449*
Análise de caminhos, 257, 297-303, *298-299*, 298n16
Análise de componentes principais (PCA), 424-433, *425-426*
  exemplos de, 426-428
  Hotteling e, 424-425

inversa, 433-434
Pearson e a, 424-425
scree-plot, 428, 430
sucesso da, 426-427
Análise de coordenadas principais (PCoA), 435-439, *440*
análise de correspondência como, 440
cálculos básicos da, 436-437
Análise de correspondência (CA), 439-444
cálculos básicos para, 440-441
destendenciada, 443-444
PCoA e, 440
RA, sinônimo de, 439
resultados de, 442
Análise de correspondência canônica (CCA), 455-456
Análise de correspondência destendenciada (DCA), 443-444
Análise de covariância (ANCOVA), 183-184
e ANOVA, 332-335, 333n6, 335
plotando os resultados da, 350-353
Análise de dados exploratória, 236
Análise de dados gráfica exploratória, 236
Análise de gradiente indireta, 439
Análise de impacto ambiental, 161-162n4. *Ver também* delineamentos ADCI
Análise de intervenção randomizada (RIA), 213-215
Análise de modelos mistos, 335
Análise de redundância (RDA), 455-462, *459-458*, *460*
básico em, 455-458
bi-plot da, 460-461
com base em distância, 461-462
exemplo de, 457-458
MANOVA *vs.*, 461-462
teste dos resultados da, 458, 460-462
variáveis ambientais usadas na, 457-458
Análise de Similaridade (ANOSIM), 410, 412-413
Análise de variância (ANOVA), 45, 221-223, 374n4
com parcelas subdivididas, 206-208, 208n7, 326-331
comparações de médias e, 351-362, 353-354n9-10, 354-355, 355-356n11, 356-357, 360-362
contrastes *a posteriori* e, 353-357, 355-356n11
contrastes *a priori* e, 353-354, 356-362
convencional, 213-214, 213n9
correlações de Bonferroni e, 362-366, 363-364n12, 364-365n13
de um fator, 342-345
delineamento proporcional e, 220-222
delineamentos ADCI e, 212-215, 212n8, 213n9, 213-214
delineamentos tabulares e, 217-221
depois da, 342-353, 349-350n8
desenvolvimento da, 135n5, 155, 307-308
dois fatores, 190-191, 201
duas vias, 321-326, 322-323n4, 324-325, 345-349
e ANCOVA, 332-335, 333n6, 335
fatores aleatórios na, 335-340
fatores fixos na, 335-340, 336-339
graus de liberdade, 374-375, 389
método, 135, 183
n-fatorial, 326
objetivo da, 307-308
partição da soma dos quadrados, 307-313, 311-312n1
partição de variâncias na, 339-342, 341-342n7
pressupostos da, 312-314
razões-F na, 137, 315-320, 316-317n2, 317-319
regra dos 30, 216-218, 217n10
réplicas na, 189-190
rótulos/símbolos na, 308
soma dos quadrados dos resíduos, 311-312
tabelas de, 268-271, 317-335
terminologia, 189-191
termos de interação e, 348-351, 349-350n8
testes de hipóteses com, 313-316
testes múltiplos e, 362-366, 363-364n12, 364-365n13
teste-t ou, 291-292n14, 405
três fatores, 326, 328-330
ver também delineamentos de ANOVA; Análise de variância multivariada (MANOVA)
Análise de variância multivariada (MANOVA), 184, 332n5, 408
Análise discriminante, 451-455, 452n12, *453*
MANOVA e, 452
soma dos quadrados e dos produtos cruzados, 452
Análise fatorial, 432-433n10
Darlingtonia, 435-436
PCA, alvos similares com, 432-433
"PCA inversa", 433-434
teste de bondade do ajuste, 435-436n11
Análises bayesianas, 77n1, 107-108n11, 122-123, 125-126, 300-301, 395-397
amostra, 125-129, 126-127n1
árvores de classificação e, 393
de tabelas de contingência, 367-368, 393-394
distribuição de probabilidades *a posteriori* em, 147-148

distribuição de probabilidades *a priori* em, 141-146, 142-143n9, 143-144n10-11
hipóteses em, 140-143, 141-142n7-8, 142-143n9, 303-305
interpretação de resultados das, 148-151
métodos, 215
objetivo, 396
parâmetros em, 142-143
passos nas, 140-152, 141-142n7-8, 142-143n9, 143-144n10-11, 145-146, 147n12-13, 148-151
pressupostos das, 150-152
regressão linear e, 283-287, 285-286n13, 295
vantagens e desvantagens das, 150-152
verossimilhança em, 147, 147n12-13
Análises de Monte Carlo, 125-126, 150-152, 286-287
amostra, 125-129, 126-127n1
distribuição nula em, 129-131, 130-131n3
passos das, 128-133, 128-129n2, 130-131n3
pressupostos das, 133
probabilidades de cauda em, 132-133
teste bicaudal em, 130-132
teste estatístico em, 129-130
teste unicaudal em 130-132
vantagens e desvantagens, 133-134, 134n4
Análises filogenéticas, 454-455
Análises Paramétricas, 125-126, 150-152
amostra, 125-129, 126-127n1
análises não paramétricas, 138-140
hipótese nula em, 137, 137n6
passos em, 135-138, 135n5, 137n6
pressupostos das, 137-139
probabilidades em, 137-138
vantagens e desvantagens, 138-139
ANOSIM, alternativa à, 410, 412-413
análises discriminantes e, 452
ANOVA análoga à, 410, 412-413
com base em distâncias RDA *vs.*, 461-462
pressupostos da, 410, 412-413
resultados de um fator, 411
ANOVA. *Ver* Análise de Variância
Antes-Depois, Controle-Impacto. *Ver* delineamentos ADCI
Anti-Stratfordianos, 98-99n4
Aphaenogaster, 320
Apostador, 29-30n8
Aranhas
lagartos e, 159-161
Linyphiidae, 63-66, 68-76, 79-80
Arcoseno/transformação do arcoseno da raiz quadrada, 248-249
ARIMA (modelo autorregressivo integrado de médias móveis), 213-215
Aristóteles, 98-99, 98-99n4
Arquivos simples, 227-228n3
Árvore lógica, 105-107n10
Árvores
de classificação, 390-393
lógicas, 105-107n10
sequoias, 161-165, 152-163n5
ASCII (*American Standard Code for Information Exchange*), 226-227n2
Assimetria, 59-89, *88*, 415
Ato de Liberdade de Informação, 225-226
Atrazina, 212
Autocorrelação, espacial, 163-164n6, *164-165*
Autovalores, 427-428, *429*, 471-475
Autovetores, 427-428, *429*, 438-439, 471-475
Avaliação de impactos, 212-213, 212n8

B

Bacon, Sir Francis, 98-99n4, 101-105
Bayes, Thomas, 42n16
Beija-flor, 168-171, *169-171*
Bernoulli, Jacob (Jacques, James), 46n2
Beta (β). *Ver* erro Tipo II
BIC. *Ver* critério de informação de Bayes
Biometrika, 373n3
Blocos, 193, 208
aleatorizados, 193-198, 195n4, 317-320
réplicas e, 193-194
Boca-de-Leão, 98n2
Bonaparte, Napoleão, 71-73n14
Bootstrap, 128-129n2, 454-455
Borboletas, 36n13
Brachymyrmex, 441

C

CA. *Ver* Análise de correspondência
Caixa preta, 111-112n13, 156n1
Cálculo, 64-65n9
Cambridge University, 34n12
Cadeira de honra de Genética, 135n5
Camponotus (formigas), 320
Cantor, Georg Ferdinand Ludwig Phillip, 25n4
Captura de presas, 24-30, 26n5, *27-29*, 28-29n6
Carga, 160-161, 426-428, 430
Caribenho, 177
Casas de Vegetação (Estufas), 174-175n11

Casinos, 29-30n8
CCA (Análise de correspondência canônica), 455-456
CD-ROMs, 229n6
Centroide, grupo, 408, 408n2
Cercados, 176
Chave taxonômica, 390
Cheaper By The Dozen (Gilbreth & Carey), 166-167n7
Chebyshev, Pafnuty, 71-73n14
Circularidade, 210-210
Classificação, 447-448
  árvores de, 390-393
  matriz de, 452, 454
  vantagens e desvantagens da, 454-456
Coeficiente binomial, 49-50, 50n4
Coeficientes
  binomiais, 49-50, 50n4
  de caminhos, 300-301
  de controle, 357-358
  de correlação produto-momento, 267-268, 268n6
  de determinação, 266-268, 268n6
  de dispersão, 91
  de variação, 90
  soma de, 356-357
Combinação linear, 426-427
Combinações, 48-49
Complemento, 36, 39
Componente principal, 425-426
  calculando o, 427-428
  escore do, 424-425, 431-432
Computadores, 133-134
  AME/S-Plus e, 134n4
  linguagens de, 134n4
  NJG / Delphi e, 134n4
  nomes de arquivos, 228
Conceito de espécie, 24n1
Conjuntos, 25
  combinando, 38-42
  conceito de, 34, 34
  discretos, 25, 25n4
  vazio, 39-40, 40n14
Conteúdo Padrão para Metadados Digitais Geoespaciais (CPMDG), 228n5
Contrastes
  a posteriori, 353-357, 355-356n11, 356-357
  a priori, 353-354, 356-362
  ortogonais, 358-359
Correção de continuidade de Yates, 379
Correlações, 184-185, 184-185n2, 189n3, 257n1
Costa do Pacífico, 81, 81n3
Covariâncias, 262-264, 264n5
Covariáveis, 179, 332
CPMDG (Conteúdo Padrão para Metadados Digitais Geoespaciais), 228n5
Crematogaster, 441
Crescimento alométrico, 238n13
Crescimento isométrico, 238n13
Critério de Informação de Akaike (AIC), 303, 305n19, 388, 393
Critério de informação de Bayes (BIC), 304, 304n18, 305n19
Curtose, 87-89, *88*, 415
Curvas
  ajuste, 121-122n22
  gaussiana, 261n4

D

D.W.E.M. (Dead White European Male), 261n3
Dados. *Ver também* Dados discrepantes e Planilhas
  angular, 248-249, 333n6
  arcoseno, transformação do arcoseno da raiz quadrada, 248-249
  basal (pré-manipulação), 213, 213n9
  binomial, 393-394
  brutos, 226-228, 226-227n2-3, 227-228n4, 228n5
  campo, 181-182, 226-227
  checagem, 230-241, 231, 232n9-10, 234n11, 235, 236n12, 237-238, 238n13, 238-240, 241n14
  compartilhar, 225-226, 225-226n1
  conjuntos, multivariados, 189, 403, 415
  curadoria/armazenagem, 228-230, 229n6-7, 230n8
  erros e, 232-233, 232n10
  faltantes, 233
  filogenéticos, 450-451
  fluxograma de manejo, 251-253
  forquilhas, dicotômica, 390
  frequência de quatro fatores, 384
  garimpagem, 236, 236n12
  GQ/CQ, 227-228n4, 241, 241n14, 250-251, 253
  homocedástico, 246-247
  independência dos, 184-185
  meta, 227-228, 227-228n4, 228n5
  modelos lineares e, 259-262, 260n2, 261n3-4, 262
  multidimensional categórico, 383
  organização, 383-386
  planilhas, 226-228, 226-227n2-3
  rastros de auditoria, 241, 241n14, 250-253
  resíduos, 246-247

transformação, 241-253, 242n15, 246n16, 246-247n17, 247-248n18, 293
transformação da raiz quadrada, 247-249
transformação de Box-Cox, 249-251
transformação inversa, 248-250
transformação logarítmica, 246-248, 246-247n17, 247-248n18
transformação reversa, 250-251
univariados, 401-402
Dados discrepantes, 90, *128-129*
dados e, 230-241, 232n9, 234n11, 236n12, 238n13
identificação de, 231-232
importância de, 230-232, 232n9
Dados multivariados, 401-402
análise fatorial e, 435-436
PCA e, 424-425
PCoA e, 435-436
propriedades de, 402
técnicas de ordenação, 424-425
Dados tabulares, 390
Darlingtonia californica (planta-jarro da Califórnia)
dados sobre, 233-241, 234n11, 235, 236n12, 237-238, 238n13, 238-240, 403, 409-411, 415, 419, 435-436
para análise fatorial, 435-436
PCA aplicada à, 426-427
regressão e, 291-293
Darwin, Charles, 24n1-2, 98n2
Darwin, Erasmus, 98n2
Days Gulch, *235*
DCA (análise de correspondência destendenciada), 443-444
De Moivre, Abraham, 67-68n11
Decis, 89
Declarações
conjunto de, 98-101
e/ou, 33
Decomposição em valores singulares, 474-475
Dedekind, Julius Wilhelm Richard, 25n4
Dedução, 98-103, 122-123
definição de, 99-101
Delineamento ADCI (Antes-Depois, Controle-Impacto), 161-162n4, 167-169
ANOVA e, 212-215, 212n8, 213n9, 213-214
de impactos ambientais, 212-215, 212n8, 213n9, 213-214
Delineamento aninhado
análise de, 199-200
ANOVA e, 198-200, 199n5
desvantagens, 200
Delineamento de blocos aleatorizados, 193-198, *194*, 195n4, *196*, *317-319*, 317-320
Delineamento de experimentos
ADCI. *Ver* delineamentos ACDI
ANOVA. *Ver* delineamentos de ANOVA
classe tabular em, 183
classes de, 183-184, 183n1
outras classes de, 183-184
regressão, 183-189, 184, 184-185n2, 189n3, 215-218, 217n10
regressão com um fator, 184-187, 184-185n2
regressão logística, 183
regressão múltipla, 185-189, 189n3
Delineamentos
aditivo, 205, *205*
amostragem, 181-182, 220-221
aninhado, 199-200
confundidos, 110-112, 158, 171-172, 174-175n11, 175
fatorial, 201, 202
medidas repetidas, 209-212, 327-332, 34n5
modelo I, II, III, 381
múltiplos fatores, *ver* plano, dois fatores
proporcional, 220-222
regressão experimental, 215-218, 217n10
substantivos, *205*, 205-207
superfície de resposta, 205-207
tabular, 217-221
três ou mais fatores, 208-209, 209n7
ver também Delineamentos de ANOVA; Delineamentos ADCI; Delineamento experimental
Delineamentos de ANOVA
aninhado, 198-200, 199n5, 320-323, 321n3
blocos aleatorizados, 193-198, 194, 195n4, 196, 317-320
efeitos aleatórios, 217n10
efeitos fixos, 217n10
medidas repetidas, 209-212, 327-332, 332n5
multifatorial/plano de dois fatores, 190, 200-207, 201n6
parcelas subdivididas, 206-208, 207-208; 208n7
plano de um fator, 191-193, 201
três ou mais fatores, 208-209
um fator, 190-193
ver também delineamentos ADCI
Delphinium muttallianum. *Ver* esporinha
Dendrograma, 449-450
Densidade de fluxo de fótons fotossintéticos (DFFF), 101-105, *102-104*

Dependência temporal, 161-164, 162-163n5, 163-164n6, *164-165*
Deslocamento, 67-69, *68-69*
Desvios, 163-164, *164-165*, 246-247
DFFF. *Ver* Densidade de fluxo de fótons fotossintéticos
Diagrama de árvore (dendrograma), 449-450
Diagramas de Venn, 34, *34*, 34n12
  de uniões/interseções, 39
  resultados dos dados em, 111-112
Diálogo de "perguntas e respostas", 232n10
Diferença (DIF), *129-131*, 129-131, 130-131n3
Diferença honestamente significante (DHS – *honestly significant difference*), *354-355*, 354-357, 355-356n11, *356-357*
Diferença honestamente significante de Tukey (DSH), *354-355*, 354-357, 355-356n11, *356-357*
Dimensão, da matriz, 465
DIP (*Dual-Inline Package*) duplo encapsulamento, 363-364n12
Disquetes, 229n6
Distância de Cook, 279-280n11
Distância de corda, *422*
Distância de Manhattan (quarteirão), *422*
Distância euclidiana, 416-417, *417-419*, 420, *421*, *422*, 431-432, 452, 454
Distâncias
  Bray-Curtis, 422
  Cook, 279-280n11
  corda, 422
  euclidiana, 416-422, 431-432, 452, 454
  Jaccard, 423n6
  Mahalanobis, 452, 454
  Manhattan, 422
  medidas comuns de, 422
  medidas de, 416-420
  métrica, 421, 423
  multivariada, 416-417
  semimétrica, 421
Distâncias de Bray Curtis, *422*
Distâncias de Jaccard, 423n6
Distâncias de Mahalanobis, 452, 454
Distribuição *a posteriori*, 143-144n10
Distribuição assimétrica, 53, *53*, *128-129*
Distribuição chi-quadrado, 374n4
Distribuição contínua, 397-399, *398-399*
Distribuição de Poisson, 55-57, 61, *61*, 397
Distribuição de probabilidades, *52*, *53*
  a posteriori, 141, 147-148
  a priori, 141-146, 142-143n8, 143-144n10-11
  funções de, 51, 54n6
  multinomial, 375-376, 375-376n5
  ver também Distribuições; Distribuição normal (gaussiana)
Distribuição de probabilidades de Gauss; *Ver* Distribuição de probabilidades normal
Distribuição gama inversa (GI), 143-144n11
Distribuição Gamma (Γ), 143-146, 143-144n11
Distribuição Leptocúrtica, 88, *89*, *94-95*
Distribuição multivariada normal, 410, 412-413, 413n3, *413-414*
  teste da, 413-417
  teste de Doornik/Hansen para, 415, 415n4
Distribuição normal (gaussiana), 65-73, 410, 412-413
  assimetria à direita, 87, 88
  assimetria à esquerda, 87-88
  contínua, 68-73, 71-72n13
  curva de sino na, 65-69, 66-67n10, 67-68n11
  deslocamento na, 67-69
  mudança de escala, 67-69
  padrão, 67-68
  propriedades da, 67-70
  unidimensional, 410, 412-413
  Ver também Distribuição de probabilidades
Distribuição normal bivariada, *403*, 413, *413*
Distribuição nula, 129-131, 130-131n3, 137, 137n6
Distribuição platicúrtica, 88, *89*
Distribuição-*t*, 93-95, *94-95*
Distribuição-*t* de Student, 93-95, 94-95n8
Distribuições, 73-74
  assimétrica, 53, 128-129
  binomial. *Ver* distribuição binomial
  Chi-quadrado, 374n4
  contínua, 397-399
  de Student, 94-95n8
  gamma, 143-146, 143-144n11
  gamma inversa, 143-144n11
  gaussiana. *Ver* Distribuição de probabilidades normal
  leptocúrtica, 88, 89, 94-95
  normal. *Ver* Distribuição de probabilidades normal
  normal bivariada, 403, 413-414
  normal multivariada, 410, 412-417
  nula, 129-131, 130-131n3, 137, 137n6
  platicúrtica, 88, 89
  Poisson, 55-57, 61, 397
  posteriori, 143-144n10
  probabilidades. *Ver* distribuição de probabilidades
  quantis, 89-90

simétrica, 52-53
t, 93-95
uniforme, 71-72, 234n10
Distribuições binomiais
gerando, 51-54
variâncias de, 48-49, 61
Doornik, Jungen A., 415
Dual-Inline Package (DIP) testes de interruptor, 363-364n12
DVDs, 229n6

E

EDA, 236
Efeito do arco, 443-444
Efeitos
aleatórios, 217n10, 325
do arco, 443-444
fixos, 217n10, 325
interação, 190-191, 325
principais, 190, 313-314, 322-323
tamanho, 165-166
tratamento, 322-326
Efeitos Aleatórios, 217n10, 325
Efeitos de interação, 190-191, 325
Efeitos fixos, 217n10, 325
Efeitos principais, 190, 313-314, 322-323
Eficiência, 217
Eixo
eixo-x, 28-29, 28-29n6, 183, 257
eixo-y, 28-29, 28-29n6, 183, 257
maior, 425-426
principal, 424-425, 424-425n9, 425-426
Elementos, de uma matriz, 403, 465
Empiricismo, 98-99n4, 101-103
EP. *Ver* Erro-padrão
Equação de Hardy-Weinberg, 40n15
Equação de Michaelis-Mentem, *102-104*, 103-104, 122-123
Erro Tipo I (α), *117*, 117-119, 118n19-20, *120*, 208
Erro Tipo II (β), *117*, 118-120, 118n20, *120*, 231
Erro-padrão (EP), 84-87, *85-87*, 85-87n6
Erro-padrão da regressão, 266
Erros
dados e, 232-233, 232n10
medidas, 162-163n5
processo, 162-163n5
ruído branco, 162-163
tipo I (α), 117-120, 118n19-20, 208
tipo II (β), 117-120, 118n20, 231
Escalonamento multidimensional não métrico (NMDS), 443-446, *446-447*
computação básica, 444
scree plot em, 445-446
Escores ajustados dos locais, 456-457
Escores das espécies, 456-457
Escores dos locais, 456-457
Esfericidade, pressuposto de, 410, 412-413
Espaço amostral, 25, 25n3
de interesse, 32
definição de, 174-175
exemplos de, 174-175, 175n12
Esponjas, 353-354n10
de fogo, 353-362
púrpura (*Haliclona implexiformis*), 353-362
Esporinha (*Delphinium nuttallianum*), 309, *310*
Essay towards solving a problem in the doctrine of chances, 42n16
Estatística, 1. *Ver também* Posição, Dispersão, medidas de
circular, 333n6
coluna, 234, 235
critério de informação, 303
descritiva, 75, 91-92, 94-95
paramétrica/frequentista/assintótica, 91-92
quatro testes, 409-411
testes de, 109-110, 129-130, 135-137
transformação de, 246-251, 246-247n17, 247-248n18
transformação de dados e, 246-251, 246-247n17, 247-248n18
Estatística chi-quadrado de Pearson, 373-374
distribuição esperada de, 374n4
tamanho amostral grande e, 378
Estatística-F
distribuição de frequências da, 461-462
quatro estatísticas do teste de transformação, 409-410
ver também razão-F
Estatísticas de critérios de informação, 303
Estimadores
M-, 287-290
sem viés, 277
Estimadores-M, 287-290, *288-289*
Estresse, cálculo, 444
Estudos observacionais, 97. *Ver também* Experimentos naturais
Euler, Leonhard, 247-248n18
Eventos, 25, 25n3
compartilhados, 33-35, 38-42, 40n14-15
complexos, 33-34, 38-42, 39, 40n14-15

mutuamente exclusivos, 41
resultados de, 32-33
união de eventos simples, 33
Evolução, 24n1, 232n9
Expectativa, *28-29*, 28-30, 59
valores, 71-72
Experimento fotográfico, 160-164
Experimentos, 181-182. *Ver também* Delineamentos Experimentais; Amostragem
casa de vegetação (estufa), 174-175n11
cercados/pilotos em, 176
competição, 205
condições ambientais em, 177
controles em, 178
covariáveis em, 179
de campo efetivos, 176-179
de pressão, 163-166
de pulso, 163-166
estudos correlativos, 97
estudos observacionais, 97
extensão/grão em, 176-177
fatores confundidos em, 171-172, 174-175n11, 175
fotográfico, 160-165, 161-162n3-4, 162-163n5, 163-164n6
independência em, 168-171
laboratório, 174-175n11
manipulativo, 97, 157-159, 160-161n2, 201-202
natural, 26, 26n5, 159-161, 160-161n2, 201-202
regra dos 10 em, 167-168
regra dos 3 em, 167-168n8
replicação em, 165-169, 166-167n7, 167-168n8, 172-176, 172-173n9-10, 174-175n11, 175n12, 178-179
trajetória, 160-165, 161-162n3-4, 162-163n5, 163-164n6
Experimentos, campo
delineando, 155
diferenças na variável Y/medidas em, 155-157
estimativa de parâmetros em, 157
ponto do, 155-157
variável Y/fator X em, 156, 156n1
Experimentos de pressão, 163-166
Experimentos de pulso, 163-166, *165-166*
Experimentos de trajetória, 160-164
Experimentos manipulativos, 97, 157-159, 160-161n2, 201-202
Experimentos naturais, 26, 26n5, 159-161, *159-160*, 160-161n2, 201-202
Extensão, 176-177, *177*
Extrapolação, 274-275

F

Faculdade Caius, 34n12
Fatores, 190
aleatório, 335-340, 336-339
carga dos, 433-434
comuns, 433-434
dentro, 331
entre-sujeitos, 209
escores, 435-436
fixos, 335-340, 336-339
intrassujeitos, 209
modelo, 435
parcela completa, 206-208, 331
rotação, 435
subparcela, 206-208
Fatores confundidos, em experimentos, 171-172, *171-172*, 174-175n11, *175*
Fatores de Bayes, 303-304, 304n18
Fatorial, 49-50, 49-50n3, 201, *202*
FCN (Fundação de Ciência Nacional, E.U.A.)
FDC (Função de distribuição cumulativa)
FDP. *Ver* Função densidade de probabilidade
Filosofia, 91-92
Fisher, Sir Ronald, 118n19, 135n5, 141-142n7, 147n13
contribuições de, 195n4, 307, 311-312
probabilidades combinadas e, 364-365, 364-365n13
Formica (formigas), 320
Formigas, 50
em parcelas pequenas, 177
Floresta, 295, 296
que forrageiam no solo, 125-129, 126-127n1, 141
Ver também Aphaenogaster; Camponotus; Formica; Myrmica; Pheidole
Fotossíntese, 101-105, *102-104*
Frequência, 28-29, *28-29*, 28-29n6
Frequências relativas, 368n1
Frequentistas, 51, 142-143n9, 147n13
FTP, 230n8
Função de distribuição cumulativa (FDC), 63-64, *63-64*, 66-67n10
Função Densidade de Probabilidade (FDP), 62-64, *63-64*, 66-67n10, 147n12
Funções, 45
distribuição, 51-53, 54n6
influência, 280-283, 282-283n12
log da verossimilhança, 249-251
monótona contínua, 242, 242n15
potência, 242, 246n16
Fundação de Ciência Nacional (FCN - E.U.A.), 161-162n3

G

Gaiola controle, 201n6
Galileo, 98-99n4
Galton, Francis, 261n3
Galton Laboratory for National Eugenics, 373n3
Garantia de qualidade/Controle de qualidade, 227-228n4, 241, 241n14, 250-251, 253
Gauss, Karl Friedrich, 67-68n11
GenBank, 225-226n1
GI (Gamma inversa), 143-144n11
gl. *Ver* Graus de liberdade
Gliszcynski, Thomasz, 29-30n7
GOPHER, 230n8
Gosset, W.S., 94-95n8
Gráfico de barras padrão, *127-128*
Gráficos
  bi-plot, 441, 460-461
  box plot, 89, 90, 128-129, 235, 237-238, 433-434
  caule-e-folhas, 235, 237-238
  completa, 206-207, *207-208*, 331
  de dispersão, 237-241, 241n13, 436-437
  dentro de, 331
  dos resíduos, 277-281, 279-280n11
  formigas, 177
  mosaico, 370, 371, 384
  scree-plot, 427-430, 445-446
  sub, 206-207, *207-208*
  subdivididas, 206-208, *207-208*, 208n7, 326-327, 330-331
Gráficos de caule-e-folha, *235*, 237-238, *238*
Gráficos de dispersão, 237-241, *238-240*, 241n13, 436-437
Gráficos de mosaicos, 370, *371*, *384*
Grãos, 176-177, *177*
Graus de liberdade (gl), 84-85, 270, 329-330, 374-375
  ANOVA e, 318-320, 374-375
  modelos log-lineares e, 388-390
  teste-G e, 377
Grubb, Howard, 29-30n7

H

Hábitats, 330
  campo/floresta, 125-129, 126-127n1, 141
  micro, 218-221
Haliclona implexiformis (esponja púrpura), 353-362, *360-362*
Hansen, Henrik, 415
Harvard Forest, 98-101, 112n14
Heisenberg, Werner, 30n9
Herança, 98n2
Herbívoros, 36n13, 326
Hidrodinâmica, 359-360
Hipótese alternativa, 110-113, 111-112n13, 112n14
Hipótese nula, 31n11, 101-105, 101-103n8, *102-104*, 314-316, 370-371
  estatística, 109-111, 116, 116n18
  euro belga e a, 395, 395
  tabelas de contingência e, 370-371
  testes de, 111-113, 112n14, 136-137
  valores de P e, 374-375
Hipóteses
  alternativas, 110-113, 111-112n13, 112n14
  ciclo de, 99-101
  científica, 98, 98n2-3, 108-109, 116, 116n18
  de trabalho, 98, 98n2-3
  definição de, 97-98, 97n1, 98n2-3
  em análises bayesianas, 140-143, 141-142n7-8, 142-143n9, 303-305
  erro Tipo I, 117-120, 118n19-20
  erro Tipo II, 117-120, 118n20
  extrínseca, 397
  intrínseca/extrínseca, 397
  metafísica, 98, 98n3
  nula. Ver Hipótese nula
  probabilidade de, 140-143, 141-142n7-8, 142-143n9
  testando, 117-120, 118n19-20
Histogramas, 28-29, *28-29*, 28-29n6
History of Mathematics (Boyer), 71-73n14
Hormônio glicocorticoide (GC), 109-110, 109-110n12, 111-114, 111-112n13, 116n18, 118-120, 120n21
Hormônios glicocorticoides, 109-114, 109-110n12, 111-112n13, 116n18, 118-120, 120n21
Hotelling, Harold, 424-425, 424-425n8
HTML, 230n8
HTTP, 230n8

I

Idade da razão, 98-99n4
Ilhas de Galápagos, 60, 242n15, *243*, *244*, *245*, 246n16, 246-247n17, 247-248n18, *252*, 260, 263-264
Ilhas do Pacífico, 273n7
Impressoras, 229, 229n6
Impureza, 390-391

Inclinação, *258*, 258-259, *259*, 268n6
Independência
  dos dados, 184-185
  em experimentos, 168-171, 181-182
  pressuposto de, 35, 36n13
Índice de Gini. *Ver* Índice de Shannon-Weiner
Índice de Shannon-Weiner, 390-391, 391n6
Índice do r^2 ajustado, 303
Indução
  definição de, 98-101, 98-99n4, 99-101n5
  desvantagens, 100-103, 101-103n7
  nos dias atuais, 101-105, 101-103n8
  vantagens da, 99-101, 99-101n5, 100-101n6
Inferência, 97, 97n1, 99-101
  bayesiana, 101-105, 101-103n8
  escopo da, 126-127
Inferência bayesiana, 101-105, 101-103n8, *102-104*
Insetos, 25n3, 27-29, 174-175n11
Inteiros
  não negativos, 49-50n3
  pares, 25n4
Intercepto, *258*, 258-259, *259*
Interpolação, 274-275
Interseção, 33, 35
Intervalos, 61
  abertos, 61-62
  de confiança. *Ver* Intervalos de confiança
  de credibilidade, 93-94, 149-151
  de predição, 274-276, 275-276n8-9
  de predição inversa, 275-276n8
  de predição simultânea, 275-276n8
  de unidade fechada, 61-62
Intervalos de confiança, *79-80*, 92-93
  de Fisher, 380
  de regressão, 271-276, 273n7, 275-276n8-9
  generalizados, 93-95
  regra dos dois DP, 93n7
  símbolo, 372
  transformação-reversa, 275-276, 275-276n9
  Ver também Intervalos

J

Jackknife, 128-129n2, 282-283, 282-283n11, 282-283n12, 454-455
João XXII (Papa), 101-103n8
Kolmogorov, Andrei, 77n1
Kuhn, Thomas, 107-108n11

L

Lagartas, 35-37, 36n13, *37*
Lagartos
  Anolis, 218-221
  aranhas e, 159-161
Lakatos, Imre, 107-108n11
Lambda de Wilk, 409-410
Laplace, Pierre, 71-73n14
Lei do Carma, 31n11
Lei dos Grande Números, 46n2, 75
  descrição da, 77-78, 77n1
  estatística paramétrica e, 91-92
  observações e, 85-87
leiaute de dois fatores, 190, 200-207, 201n6, *202-205*
  fatores dos tratamentos em, 200-203, 201n6
  temas de, 205-207
  vantagens e desvantagens203-205
Linguagem de Metadados Ecológicos (LME), 228n5
Linha reta, 257-259, 257n1, *258*, *259*, 260n2
Linyphiidae. Ver Aranhas
Littoraria angulifera (molusco), 447-448
LME (Linguagem de Metadados Ecológicos), 228n5
Lobos, efeitos de carros de neve sobre, 109-110, 109-110n12, 110-113, 111-112n13, 112n14
Logaritmo, natural, 247-248n18
Luz, 101-105, *102-104*
Lyapounov, Alexander, 71-73n14
LYNX, 230n8

M

Maior raiz de Roy, 409-410
MANOVA. *Ver* Análise de variância multivariada
Markov, Andrei, 71-73n14
Marx, Karl, 104-105n9
MatLab, 66-67n10
Matriz, 403, 463-463
  binária de presença e ausência, 435-436, 438
  de classificação, 452, 454
  de comunidade, 264n5
  de distância, 438
  de distância euclidiana, 449-450
  de variância e covariância, 264n5, 404, 406
  diagonal, 404-405, 468
  E/H, 409-411
  formas quadráticas, 469-471
  H/E, 409-411
  identidade, 468

inversa, 469-470
local × espécies, 420
ortogonal, 469-470
ortonormal, 441
quadrada, 404, 465-466
sagaz, 282-283n12
simétrica, 468
singular, 473-475
ver também Soma dos quadrados e dos produtos cruzados
Matriz quadrada, 465-466
MC (Momento Central), 87-89
McArthur, Robert H., 100-101n6
Mecânica Celestial (Mécanique celeste, Laplace), 71-73n14
Mécanique céleste [Mecânica Celéste] (Laplace), 71-73n14
Média
agrupamento K-, 451
ajustada, 333-334
aritmética, 76-78
comparação de, 109-110, 109-110n12
EP da, 84-87
geométrica, 79-81
harmônica, 81, 81n3
outra posição, 78-81, 78n2
transformação inversa, 79-80
vetor de, 403
Média aritmética, 76-78
Média harmônica, 81
Mediana, *82*, 82-83
Médias recíprocas
análise de gradiente indireto, sinônimo de, 439
CA, sinônimo de, 439
Medida de dissimilaridade de Sørensen, 438, 445-446
Medidas, 23, 46n1
Medidas de dispersão
assimetria em, 87-89
curtose e, 87-89
desvio-padrão, 83-87, 84-85n5
EP da média como, 84-87, 85-87n6
MC e, 87-89
quantis, 89-90
uso de, 90-91
variância, 83-87
Medidas de dissimilaridade, *439*
Mercator, Gerardus, 247-248n18
Mercator, Nicolaus, 247-248n18
Meta-análises, 341-342n7
Metadados, 227-228, 227-228n4, 228n5
Método de mínimos quadrados, 257, *296*
Método hipotético-dedutivo, 122-123
desenvolvimento do, 104-107, 104-105n9, 115n17
desvantagens, 105-108, 107-108n11
vantagens, 105-107, 105-107n10, 115n17
Métodos científicos, 98, 98n2-3, 122-123
dedução como, 98-103
hipotético-dedutivo, 104-108, 104-105n9, 105-107n10, 107-108n11, 115n17
indução como, 98-103, 98-99n4, 99-101n5
Métodos de agrupamento hierárquico, 449-451
Métodos de agrupamento não hierárquicos, 449-451
Métodos de Bonferroni, 362-366, 363-364n12, 364-365n13
Métodos de Dunn-Sidak, 362-365
Métodos de eliminação regressiva, 301
Métodos de seleção progressiva, 301
Métodos gradativos, 301
Mexilhão, *194*, 194-195, 197. *Ver também* Análise de variância
estudo de, 209, 210
moluscos, 199n5, 200-203, 201n6
Microhábitats, 218-221
Mínimos quadrados aparados, 287-290, *288-289*
Mirifici logarithmorum canonis descriptio (Napier), 246-247n17, 247-248n18
Modelo autorregressivo integrado de médias móveis (ARIMA), 213-215
Modelo de colonização passiva, *298-299*
Modelos autorregressivos, 163-164
Modelos de séries temporais, 162-163, 162-163n5
Modelos hierárquicos, 388-390
Modelos lineares, 259-262, 260n2, 261n3-4, *262*
Modelos log-lineares386
gl para, 388-390
hierárquico, 387, 388
métodos bayesianos, 393
Moedas
arremesso, 29-31, 29-30n7-8, 30n9-10, 31n11, 54n6
dados e, 172-173n10
euro, 29-30n7, 54n6
euro belga, 29-30n7, 395, 395
honestas, 54n6
viciadas, 29-30n8
Moluscos
Littoraria angulifera, 447-448
mexilhão, 200-202, 201n6, 202, 203

Momento central (MC), 87-89
Montanhas Siskiyou, 234n11, 235
Monte Carlo, 128-129n2, 381-383
Mudança de escala, 67-69, *68-69*
Multicolinearidade, 189, 296
Museu de História computacional, 229n6
Myrmica (formigas), 23-24, 112n14, 320

N

Napier, John 246-247n17, 247-248n18
Navalha de Ockham, 101-103n8
Neutrino, 101-103n7
Newton, Sir Isaac, 98-99n4, 104-105
Novum organum (Bacon), 98-99n4
Números irracionais, 25n4
Números racionais, 25n4

O

O Arco-íris da Gravidade (Gravity's Rainbow), 55n7
Observações, 99-103, *100-101*, 181-182
Ockham, Sir William, 101-103n8
Odocoileus hemionus. *Ver* veado-mula
Ordenação
  alvo da, 447-448
  análise de correspondência e, 442
  métodos de, 445-446
  tipos de, 424-425
  vantagens/desvantagens, 445-446
Ordenada, 183
Oregon, 234n11, *235*
Outliers. *Ver* Dados discrepantes

P

P de Fisher, 364-365n13
Padrões, 110-111, 115, 115n16
Padrões de Informação e Classificação, 228n5
Paineirinha, 35-37, 36n13, *37*
Paradigma bayesano, 42
Paradigmas, 107-108n11
Parâmetros, 66-67n10, 75
  ao nível de população, 91
  em análises bayesianas, 142-143
  estimativa, 116, *121-122*, 121-123, 121-122n22, 157
  estimativa de, mínimos quadrados, 264-266, 264n5
  predição, 121-123, 121-122n22
  regressão, 189, 189n3, 264
  regressão parcial, 294
  regressão/linha reta e, 257-259, 257n1, 260n2
  taxa, 55
Paratrechina, 441
Pauli, Wolfgang, 101-103n7
PCA, *Ver* Análise de componentes principais
PCA inversa, 433-434
PCoA- *Ver* Análise de coordenadas principais
Pearson, Karl, 67-68n11, 118n9, 364-365n13, 373n3, 424-425, 424-425n7
Peixe, 50n4, 201n6
Percentil, 89
Permutações, 50, 50n4
Pesquisa Ecológica de Longa Duração (PELD), 161-162n3
Pheidole (formigas), 320
Philosophical Transactions (Bayes), 42n16
Pitágoras de Samos, 416-417n5
Planilhas, 126-127, 126-127n1
  célula de, 226-227
  dados e, 226-228, 226-227n2-3
  programas, 226-227n2, 261n4
Plano
  pares casados, 196-197
  um fator, 191-193
Plano, dois fatores, 190, 200-207, 201n6, *202-205*
  assuntos em, 205-207
  fatores dos tratamentos em, 200-203, 201n6
  vantagens e desvantagens, 203-205
Planta-jarro. *Ver Darlingtonia californica*
Planta-jarro da Califórnia. *Ver Darlingtonia californica*
Plantas
  carnívoras, 24-30, 26n5, 27-29, 28-29n6, 31
  censo, 177
  crescimento de plantas alpinas, 309
  estudo da raridade, 47-50, 56-58
  fisiologia, 102-104
  jarro. *Ver* Darlingtonia californica; Sarracenia purpurea
  mangue-vermelho (*Rhizophora mangle*), 353-354, 353-354n9, 354-357, 360-362
  nitrogênio e, 190-191
  Rhexia mariana, 47-50, 56-58
Platão, 104-105n9
Poder $(1 - \beta)$, estatístico, 118-120, *120*
Poder estatístico $(1 - \beta)$, 118-120, *120*
Poisson, Siméon-Denis, 55n7
Poluentes, 118n20
Ponera, 441

Popper, Karl, 104-105, 104-105n9, 107-108n11
Populações
  amostral, 126-127
  crescimento de, 78-81, 78n2, 81n3
  densidade de, 338-339
  genéticas, 40n15
  isoladas, 232n9
  tamanho efetivo de, 81, 81n3
Posição, 75
  aleatória, 125-127
  média aritmética e, 76-77
  média geométrica, 79-81
  mediana/moda, 82-83
  medidas de, 76-83, 83n4
  outras médias, 78-81, 78n2
Precisão, 61-62, 234-236, *235*
Predadores, 187-189, 326
Predição, 99-101, *100-101*
  intervalos de, 274-276, 275-276n8-9
  intervalos de, inversa, 275-276n8
  parâmetros de, 121-123, 121-122n22
Primeiro Axioma da Probabilidade, 32-33
  exemplos, 47-48, 141, 147n12
Princípio da Incerteza de Heisenberg, 30n9
Princípio de Parcimônia, 101-103n8
Probabilidade, 21 *Ver também* Distribuição de probabilidades; Hipóteses
  a priori, 43
  cálculo, 44
  cálculos de, 35-37, 36n13, 37
  combinadas, 364-365, 364-365n13
  condicionais, 41-44
  de cauda. *Ver* Probabilidades de cauda
  definição de, 24, 29-31, 29-30n7-8, 30n9-10, 31n11
  estimativa de, 27-30, 28-29n6
  exata, 52n5, 132
  funções de distribuição de, 51-53, 54n6
  introdução à, 23-24, 24n1
  inversa, 141-142n7
  matemática da, 31-42, 34, 34n12, 37, 39, 40n14-15
  medindo, 24-30, 24n2, 25n3-4, 26n5, 28-29n6
  primeiro axioma, 32-33, 47-48, 141, 147n12
  segundo axioma, 34-35, 48-49
Probabilidades de cauda, 52n5, 63-64, 137-138, *137-138*
  em análises de Monte Carlo, 132-133
  em análises paramétricas, 137-138
  em regressão, 270-271
Processo de Markov, 77n1
Programa
  estatísticos, 99-101n5, 126-127n1
  pacotes, 134n4
  planilhas, 226-227n2, 261n4
Programas de pesquisas científicas (PPCs), 107-108n11
Pynchon, Thomas, 55n7

Q

QM. *Ver* Média
Quadrados
  latinos, 195n4
  médios, 83-84, 270, 314-316, 321, 321n3
  método de mínimos, 257
  mínimos, aparados, 287-290
  soma dos (SQ), 83-84, 270
Quadrados latinos, 195n4
Quantificação, 23
Quantis
  distribuições, 89-90
  regressão, 257, 289-292, *290-291*
Quartis, 89-90
Quincunx, 261n3

R

RA. *Ver* Médias recíprocas
Razão
  de chances *a posteriori*, 304, 304n18
  de chances *a priori*, 304
  Ver também Razão-F
Razão de chances *a posteriori*, 304, 304n18
Razão de chances *a priori*, 304
Razão-F, 137, 137n6
  distribuição, 137-138
  na ANOVA, 137, 315-320, 316-317n2
  regressão e, 270-271
  valor de, 141-142, 141-142n8
  ver também estatística-F
RDA. *Ver* Análise de redundância
Recherches sur La probabilité dès judgments (Poisson), 55n7
Reconstrução filogenética, 232n9
Regra dos 10, 167-168, 189
  ANOVA e, 216-218, 217n10
Regra dos 5, 167-168n8
Regressão, 155
  análise de caminhos na, 257, 297-299, 298n16, 302-303
  análise de Monte Carlo e, 282-285

análises bayesianas e, 283-287, 285-286n13
análises de, 286-301
árvores de, 393
coeficiente de determinação em, 266-268, 268n6
componentes de variância na, 266-268, 268n6
covariâncias/variâncias em, 262-264, 264n5
critérios de seleção de modelos na, 300-305, 301n17
definição de, 257
em delineamentos amostrais, 183-189, 184-185n2, 189n3, 215-218, 217n10
erro-padrão da, 266
estimativa de parâmetros por mínimos quadrados na, 264-266, 264n5
experimental, 215-218, 217n10
gl com, 374-375
inclinação da, 268n6
intervalos de confiança, 271-276, 273n7, 275-276n8-9
linear, 283-287 285-286n13, 295
linha reta/parâmetros e, 257-259, 257n1
logística, 257, 291-293, 291-292n14, 293
melhor ajuste, 261
modelos lineares na, 259-262, 260n2, 261n3-4
múltipla, 185-189, 189n3, 293-297, 301-302, 301n17
múltipla multivariada, 455-462
não linear, 257, 293
outros testes em, 271-276
parâmetros da, 189, 189n3, 264, 294
pressupostos da, 275-277, 276-277n10
probabilidades de cauda, 270-271
quantílica, 257, 289-292
razão-F e, 270-271
robusta, 257, 286-290
SQR e, 286-287
testes de hipóteses com, 268-276, 273n7, 275-276n8-9
testes diagnósticos, 277-283, 278-282, 279-280n11
um fator, 184-187, 184-185n2
Regression towards mediocrity in hereditary structure, 261n3
Reificação, 246n16
Replicação
adequada, 184-185
em experimentos, 165-169, 166-167n7, 167-168n8, 172-176, 172-173n9-10, 174-175n11, 175n12, 178-179
regra dos 30, 167-168, 189
Réplicas, 26
acessíveis, 166-167, 166-167n7
atribuição de, 193-196, 195n4, 196
blocos e, 193-194
covariáveis em, 179
em ANOVA, 189-190
espacial, 213
número de, 198-200
Resíduos, 260-261, 313-314
dados, 246-247
gráfico, 277-281, 279-280n11
normalizados, 279-280n11
padronizados, 279-280n11
variação dos, 311-312
Resultados
completos, 32-33
discretos, 25
mutuamente exclusivos, 32-33
Rhexia mariana, 47-50, 56-58, *57-58*
Rhizophora (mangue, folhas), *102-104*, 102-104
Rhizophora mangle (mangue vermelho), 353-354, 353-354n9, *354-357*
RIA (Análise de intervenção randomizada), 213-215
Roedores, deserto, *184*, 184-187, 184-185n2, *185-187*
Rotação oblíqua, 435
Rotação ortogonal, 435
Rotação varimax, 435
Royal Society of London, 42n16
Ruído branco, 162-163, 162-163n5
Ruído vermelho, 162-163n5

S

Salinidade, 182-183
Sapos, 212
Sarracenia purpúrea (plantas-jarro), 24-30, 26n5, *27-29*, 28-29n6
teia alimentar, 208-209, 208n7
Schrödinger, Erwin, 30n9
Scree-plots, 427-428, *428-430*, *445-446*
Sedativos, 349-350n8
Segundo Axioma da Probabilidade, 34-35, 48-49
Seleção natural, 24n1, 98n2
Serra Nevada, 160-161n2, 234n11
Shakespeare, William, 98-99n4
Silogismo, 98-101
Sinal de somatório, 32
Sistema de Informação Geográfica (SIG), 228n5
Solo, 228n5
Soma dos produtos cruzados, 263-264
Soma dos quadrados, 80, 83-84, 261n4, 262-264, 270. *Ver também* Soma dos quadrados dos resíduos
do erro, 318-319
partição da, 307-313, 311-312n1

Soma dos quadrados dos resíduos, 261-262, *262*, 265, 266
  ANOVA, 311-312
  regressão e, 286-287
Soma dos quadrados e produtos cruzados (SQPC), 409
  análise discriminante e, 452
  entre grupos/dentro de grupos, 409, 411
Spearman, Charles, 432-433n10
S-Plus, 66-67n10, 386
SQPC. *Ver* Soma dos quadrados e produtos cruzados
SQR. *Ver* Soma dos quadrados dos resíduos
Subconjunto, próprio, 34, *34*

T

Tabelas
  ANOVA, 268-271, 317-335
  dois fatores, 383
  L × C, 377
Tabelas de contingência
  análises, 218-219
  análises bayesianas, 367-368, 393-394
  delineamento de diferenças e, 381
  equações para, 371, 372
  espécies de plantas raras, 369
  estabelecendo, 369
  hipótese nula e, 370-371
  multifatores, 383-394
  quantidades derivadas de, 370
  quatro fatores, 383
  teste de chi-quadrado, 372
  variações entre pacotes estatísticos e, 370n2
Tabelas de contingência, dois fatores
  chi-quadrado / teste-G, 377
  dados categóricos, 363
  dados organizados em, 368-370
  frequências, 368, 368n1
  múltipla, 383-383
  proteção *vs. status* da população, 369
  valores esperados em, 383-386
  vs. tabelas multifatoriais, 383
Tabelas de contingência de dois fatores. *Ver* Tabelas de contingência, dois fatores
Tabelas L × C, 108-110, 109-110n12
  cálculo do valor de P para, 379
  chi-quadrado / teste-G para, 377-381
Tamanho dos efeitos, 165-166
Taxa de erro por experimento, 362-363
Tedania ignis (esponja de fogo), 353-362, *360-362*
Tentativas, 26
Tentativas de Bernoulli, 47-50
Teorema de Bayes, 21-44, 42n16, 143-144n10
  métodos, 303-304
  modificações do, 141, 283-285
Teorema de Pitágoras, 410, 412-413n1, 416-417, 416-417n5
Teorema de probabilidade inversa. *Ver* Teorema de Bayes
Teorema do Limite Central
  apelo ao, 162-163n5
  uso do, 83, 211-212
Teorema Fundamental da Álgebra, 67-68n11
Teorema Fundamental da Aritmética, 67-68n11
Teoria, 25n4
  da evolução, 24n1
Termos de interação, 348-351, 349-350n8
Tertuliano/tertuliana, 98n3
Teste de chi-quadrado de Pearson, 381
Teste de razões de verossimilhança. *Ver* Teste G
Teste exato de Fisher, 367-368, 379, 380
Teste T^2 de Hotteling, 405-408
Teste-G, 367-368, 393-394
  cálculo, 373, 375-376
  distribuição de probabilidade multinomial e, 375-376
  gl par, 377
  modelo de delineamento I, II, III, 381
  para tabelas L × C, 377-381
Testes
  assintóticos, 379
  bondade do ajuste, 368, 393-394, 397-398, 435-436n11
  de chi-quadrado, 367-368, 372-377
  de chi-quadrado Pearson, 381
  de diferença honestamente significante de Tukey (DHS), 354-357, 355-356n11
  de normalidade multivariada, 413-417
  Doornik/Hansen, 415, 415n4
  escolha dos, 381-383
  estatísticos, 375-376
  -G, 367-368, 377, 381
  Kolmogorov-Smirnov, 368
  múltiplos, 362-366, 363-364n12, 364-365n13
  não paramétricos, 393
  paramétricos, 393
  -t, 291-292n14, 405
  T de Hotelling, 406
  Teste de interruptor com duplo encapsulamento (DIP – *Dual-Inline Package*), 363-364n12
  teste exato de Fisher, 367-368, 379, 380

Testes assintóticos, 379
Testes de bondade do ajuste, 368, 393-394, 397-398, 435-436n11
  análise fatorial, 435-436n11
  distribuições contínuas e, 397-398
Testes de chi-quadrado, 367-368, 372-377
Testes de Doornik-Hansen, 415, 415n4
Testes de Kolmogorov-Smirnov, 368, 397-399, 415
Testes estatísticos aplicados a, 369
Testes paramétricos, 393
Teste-*t*, 291-292n14, 405
The design of experiments (Fisher)135n5
The Doctrine of Chances (Gauss), 67-68n11
The ecological Detective (Hilborn & Mangel), 134n4
The Laws of Organic Life (Darwin), 98n2
The Logic of Scientific Discovery (Popper), 104-105n9
The Open Society and Its Enemies (Popper), 104-105n9
The origin of Species (Darwin), 98n2
The structure of Scientific Revolutions (Kuhn), 107-108n11
Tolerância, 302
Total das margens, 218-221, *219-220*
Traço de Hotelling-Lawley, 409-410
Traço do Pillai, 409-410, 412-413
Traité de mécanique (Poisson), 55
Transformação angular, 248-249
Transformação da raiz quadrada, 247-249
Transformação de Box-Cox, 249-251
Transformação logarítmica (transformação log), 246-248, 246-247n17, 247-248n18
Transformação recíproca, ou inversa, 248-250
Transformações
  angular, 248-249
  arcoseno/arcoseno da raiz quadrada, 248-249
  Box-Cox, 249-251
  de dados, 241-253, 242n15, 246n16, 246-247n17, 247-248n18, 293
  estatísticas/dados, 246-251, 246-247n17, 247-248n18
  logarítmica, 246-248, 246-247n17, 247-248n18
  quatro estatísticas teste, 409-410
  raiz quadrada, 247-249
  recíproca ou inversa, 248-250
Tratamentos
  atribuição de, 213
  completamente cruzados/ortogonais, 201, 202
  efeitos dos, 322-326
  fatores de, 200-203, 201n6, 208
  níveis, 190
Trutat, 160-161n2
TWINSPAN (análise de espécies indicadoras com dois fatores), 449-450

U

Usina de energia nuclear, 212, 212n8

V

Valores. *Ver também* Valores de probabilidades
  alavancagem, 279-280n11
  de jackknife, 279-280n11
  esperados, 383-386
  fora da diagonal, 452, 454
  precisão dos, 236
  preditores, 185-189
Valores das colunas, 234-236, *235*
Valores de *P*. *Ver* Valores de probabilidades
Valores de probabilidade, 111-115
  calculando, 374-377
  determinação de, 113, 122-123
  distribuições e, 71-73
  exemplos de, 111-113, 112n14
  hipótese nula e, 374-375
  pequenos, 113-115, 114n15, 115n16-17
  significância estatística e, 108-110, 109-110n12
  tabelas L × C, 108-110, 109-110n12
  teste de chi-quadrado e, 374-376
  total fixo da linhas/colunas e, 380
Valores esperados, 59
  cálculo, 371-372
  de variáveis aleatórias contínuas, 64-66, 64-65n9
  para tabelas de contingência de dois fatores, 383-386
  pequenos, 378
Valores fora da diagonal, 452, 454
Variação, 24, 59
  componentes de, 309-313
  dentro de grupos, 311
  entre grupos, 309-310
  erro, 311-312
  residual, 311-312
Variância, 60, 60n8, 262-264, 264n5, 276-277
  componentes, 266-268, 268n6
  heterocedastica, 278-280
  homogêneas, 266-268, 268n6, 312-314

matriz de covariância, 404, 409
símbolo, padrão, 308
Variáveis, 276-277
categóricas, 182-183, 222-223, 367-368
colineares, 189
contínuas, 182-183, 222-223
dependentes, 183, 221-223
escolha de, 388
independentes, 183, 221-223, 383, 386
ordinais/ranqueadas, 182-183
preditores múltiplos, 367-368
resposta, 102-103, 183-185, 257, 367-368
três preditores, 387
Variáveis, preditoras, 102-103, 183, 183n1
categórica única, 367-368
descrição de, 189n3, 257
distribuição de, 185-187
níveis de, 184-185
ortogonais, 296-297
únicas, 295-297
Variáveis aleatórias, 45, 73-74, 142-143
binomial, 48-49, 61
de Bernoulli, 61
de Poisson, 54-58, 55n7, 61
uniforme, 71-72
variâncias para, 60, 60n8
Variáveis aleatórias, contínuas, 45-46, 61-72, *71-72*
definição de, 182-183
exponencial, 71-72, 71-72n13
log-normal, 68-71
normal, 65-66, 66-67n10, 67-68n11, 68-73, 71-72n13
uniforme, 61-64, 71-72
valores esperados, 64-66, 64-65n9
variáveis contínuas como, 182-183
Variáveis aleatórias, discretas, 45-46
definição de, 182-183
variância de, 59-61
Variáveis aleatórias, normais, *65-66*, 66-67n10
log-, 68-72
padrão, 67-68, 67-68n11
propriedades de, 67-72, 67-68n11
Variáveis aleatórias, Poisson
descrição das, 54-57
exemplos de, 56-58
família de, 56-57, 71-72, *71-72*
variáveis para, 61, *61*
Variáveis aleatórias contínuas, 45-46, 61-72
definição de, 61, 182-183
exponencial, 71-72
log-normal, 68-72
normal (gaussiana), 65-70
uniforme, 61-64, 71-72
valores esperados de, 64-66, 64-65n9
variáveis contínuas, 182-183
Variáveis aleatórias de Poisson, 54-58
Variáveis aleatórias discretas, 45-46
Bernoulli, 46-48
binomial, 48-50
definição de, 182-183
Poisson, 54-58
variância de, 59-61
Variável aleatória binomial, 48-50
Variável aleatória de Bernoulli, 46-48, 61, *61*
Variável aleatória de Gauss. *Ver* Variável aleatória, normal
Veado. *Ver* Veado mula
Veado mula (*Odocoileus hemionus*), 78-80, 78n2
taxa populacional de, 80-81, 81n3
Venn, John, 34n12
Verossimilhança, 141-142, 141-142n7
cálculo da, 147, 147n12-13
Vetor de médias, 403
Vetor linha, 403, 465-466
Vetores, Matriz, 465-466
Vetores coluna, 405, 465-466

W

World Wild Web, 230, 230n8
Wright, Sewall, 298n16

X

XML, 230n8

Z

Zoonomia (Darwin), 98n2
Z-score, 418-419
Zwadowski, Waclaw, 29-30n7